Estimating a timescale for the tree of life using integrated fossil and genomic methods

Holly Catriona Betts

A dissertation submitted to the University of Bristol in accordance with the requirements for award of the degree of Doctor of Philosophy in the Faculty of Science

School of Earth Sciences

January 2020

Word count: 27795

Abstract

Palaeontologists have traditionally tried to date the tree of life using the fossil record, which is patchy owing to the differentially preserved rock layers, the reworking of these layers by geological processes and the varying fossilization potentials of the organisms. These problems become especially acute for the most ancient scions in the tree of life. However, new fossils found in the Archaean and Hadean are constantly being described and reinterpreted, leading to a fluctuating timescale which does not allow for the analysis of evolutionary hypotheses. This thesis approaches this issue by synthesising knowledge from the fossil record and molecular dating strategies and produces a robust timescale for the tree of life. I used genetic data from extant organisms in the Eubacteria, Archaebacteria and Eukaryota. Dating the origin of these lineages and the last universal common ancestor (LUCA) required application of methods over greater timescales than they are normally applied to. It also includes the use of recently developed approaches, such as the fossilized birth death process and cross-bracing, where restrictions are applied such that multiple nodes of a tree simultaneously evolve, as in the same speciation event in a dated gene tree. The combined approach of fossil calibrations can molecular clock methodology dates the origin of lineages more accurately. The results of this thesis show that life shares a common ancestor with an age close to the that of the Earth. A lag follows before the origin of the two domains, Archaebacteria and Eubacteria in the Archaean. Crown eukaryotes appear much later in the Proterozoic. These ages show that life evolves prior to the oldest known fossils. As reconstructed timelines improve, they can help us to elucidate more about the evolution of life and the changing world which it both inhabits and influences.

Author’s Declaration

I declare that the work in this dissertation was carried out in accordance with the requirements of the University's *Regulations and Code of Practice for Research Degree Programmes* and that it has not been submitted for any other academic award. Except where indicated by specific reference in the text, the work is the candidate's own work. Work done in collaboration with, or with the assistance of, others, is indicated as such. Any views expressed in the dissertation are those of the author.

Signed: H.C. Betts Date: 14/01/2020

Acknowledgements

There are many people who have helped me to produce this thesis and to whom I owe a great deal of thanks. First and foremost, I would like to recognise the help that my supervisors, Davide Pisani and Philip Donoghue gave to me, starting during my undergraduate degree and persisting throughout my PhD. Invaluable advice was also provided by Tom Williams who was co-opted onto my supervisory team. They have all contributed so much to my understanding of this topic and have patiently helped me to produce this work and thesis.

I would also like to thank my collaborators Mark Puttick, Joe O'Reilly and James Clark for their insight and knowledge which stretches above and beyond my own. They have helped me through the nitty - gritty parts of some analyses. In addition to this the people who I have emailed in order to plunder their phylogenetic knowledge, chiefly Mario dos Reis and Ziheng Yang. In this section I would also like to thank Emily Rayfield for her help in clarifying my thought and work processes during my annual progress meetings.

I would like to acknowledge others in the department who have also been on this journey and who have provided people with which to sip tea and exchange or mitigate worries. I would like to extend my particular thanks to Frankie, AJ and James, Bex and Claire. Additionally, to Celine for taking the time to help me especially in the last stages of creating and defending my thesis. My family have been wonderful during this whole process and have really provided a calm and relaxed place for me to retreat to when necessary. I am extremely grateful for their support. In this vein I would also like to thank the support from my extra-departmental friends, who have provided a welcome change and constant support when needed; Rachel, Zoë, Ben, Charlie, Kathryn, Jess, Eve and Joss.

Contents

List of Figures

List of Tables

Thesis outline

The aim of this thesis was to produce a robust timescale for the tree of life with a focus on the most ancient nodes. In general, these nodes, such as the last universal common ancestor, have the least fossil material available with which to estimate their divergence. Thus, they are therefore in most need of a molecular clock approach which integrates fossil data with information from modern organisms. However, the oldest nodes in the tree of life have been least studied using this methodology even though it is perhaps where it could be most beneficial. This thesis has employed such methods to date a tree of life.

Chapter 1 provides an overview to the field of molecular clock methodology in which I summarise the approaches to producing a dated tree of life. This includes a brief introduction to the fossil record and the problems with using it as the only information. I give an overview of how our understanding of the tree of life has changed over time. This tree has been updated as we have become aware of more organisms and as new methods have been introduced in order to study them. What we now know of its underlying framework is important to dating it. I also introduce the molecular clock as well as the competing molecular clock calibrations strategies used in this thesis.

In Chapter 2 I outline a node calibration approach to dating the tree of life. This first involved careful assessment of the fossil record in order to produce calibrations, laid out in the results section. These calibrations are conservative in their nature and have been constructed using guidelines in order to stand up to scrutiny. In actually using the molecular clock we interrogated the various parameters which could be changed in order to model the way in which the calibrations are used. This chapter is published as ‘Betts et al., 2018; Integrated genomic and

fossil evidence illuminates life's early evolution and eukaryote origin' and illustrates that a combined approach, the utilisation of fossils and molecular data, produces a robust timescale for the tree of life.

Chapter 3 explores a solution for finding a date for the node at the root of the tree of life, the last universal common ancestor. This node is the most difficult to date because of the paucity of information. In this chapter I used genes that have a duplication event prior to LUCA to try and elucidate its age. These genes have previously been used to find a root for the tree for life and the duplication events can be leveraged to provide more powder to date LUCA by providing a node above it. This chapter uses the same calibrations as detailed in Chapter 2 but applies them on both sides of the duplication. This involves the application of two kinds of methodology. Cross-bracing and cross-calibration, both of which utilize the combined information from speciation events on either side of the duplicated node.

Chapter 4 takes a detailed look at the possible divergence time estimation of the eukaryote lineage using Bayesian phylogenetics and the fossilized birth-death model. This involved using more fossil data than in the previous chapters and thus a sample of the extensive selection of material available in the Palaeobiology Database. This methodology in theory allows the use of this extra fossil data without constraining it in a strict manner and making it part of the evolutionary process. However, it can provide misleading estimates of the divergence time if certain parameters are not well accounted for. If the parameters are used appropriately, especially the rate of fossil sampling, it can produce reasonable age estimates. Here we use it to show that eukaryotes evolved with the Palaeo-Mesoproterozoic.

Chapters 3 and 4 will be submitted for publication at the earliest possible time.

Chapter 1

Introduction: Divergence time estimation for the tree of life

Author contributions: This chapter was written and developed by H.C. Betts. Comments were provided on a draft by T.A. Williams, P.C.J. Donoghue and D. Pisani. H.C.B. contributed to ~97% of the work in this chapter.

1.1 Estimating a time tree of life

Deriving a timescale for the tree of life is something that has been attempted by palaeontologists through use of the fossil record, and, more recently, by molecular biologists who have sought to unite genetic data from extant taxa, with temporal information from fossils, in order to produce a comprehensive timeline for life. The first billion years of Earth's history set the stage for life. However, we know little about life during this time period due to the lack of available rock and thus the lack of fossils. Despite the dearth of available records from the Hadean (4520-4000 Ma), Archaean (4000-2500 Ma) and Palaeoproterozoic (2500–1600 Ma) a very literal reading of the fossil record has held sway across these periods. Each new fossil discovery and reinterpretation has caused the timescale to shift resulting in an inconsistent record from which little can be garnered about the co-evolution of life and Earth, and, more specifically, about the temporal existence of the last universal common ancestor (LUCA) and the last eukaryotic common ancestor (LECA). Both of these 'organisms' represent fundamental transitions in the history of life about which their timing can be informative. In the case of LECA we can also look at this transition with respect to the first eukaryotic common ancestor (FECA) and how long eukaryotes took to gain all their fundamental characteristics, for example the mitochondria (Pittis and Gabaldón 2016).

The fundamental tool used to estimate divergence times is the "Molecular clock". Originally introduced and applied as a rather simplistic concept by Zuckerkandl and Pauling almost 50 years ago (Zuckerkandl and Pauling 1965, 1962), the molecular clock has recently been developed into a powerful probabilistic tool (Thorne, Kishino, and Painter 1998; Sanderson 1997; Ronquist, Klopfstein, et al. 2012; Drummond et al. 2006; Heath, Huelsenbeck, and Stadler 2014; Yang and Rannala 2006). The clock allows the inference of "*the rate of evolution of the rate of molecular evolution*" (Thorne, Kishino, and Painter 1998) and can subsequently be applied to the deduction of divergence times exploiting fossils and genomic data in all those cases (a large majority) where molecular data did not evolve following a "strict" (i.e. a constant rate of substitution across sites and lineages) clock. Using these modern "relaxed" molecular clock models it is now possible to integrate data from fossil and extant taxa thus

utilising the wealth of information locked in the genes and morphology of modern organisms to date ancient events for which there is a paucity of fossil data. Importantly, modern (relaxed) implementations of the molecular clock make use of Bayesian statistics (Thorne, Kishino, and Painter 1998). The use of a Bayesian framework helps to account for the uncertainty involved in using fossil date information and disentangles rate and time from estimates of branch lengths in phylogenetic trees. Such integrative timescales can then be compared to the geological record and used to evaluate hypotheses on the co-evolution of life and Earth. This thesis establishes the use of an integrated approach where genomic and fossil data are combined to generate a timescale for ancient divergences in the tree of life.

1.2 Why date a tree of life?

Studying the timescale of the tree of life is important because it can help us to try to understand whether life was influenced by large scale geological changes or whether it helped these changes to occur. Earth emerged just over 4.53 billion years ago as a slowly aggregating ball of rock, gases and liquids with its final formation period beginning with the moon forming impact ~4.52 billion years ago. Life likely evolved sometime in the Archaean given what we know from the fossils (Sugitani et al. 2013; Sugitani, Mimura, Takeuchi, Lepot, et al. 2015; Wacey et al. 2011) and has survived large scale changes ever since. The Great Oxidation Event (GOE) was the first major appreciable change in oxygen levels in Earth's history. In theory, the rise in oxygen levels could have been caused by the evolution of Cyanobacteria, a lineage of Eubacteria capable of oxygen production via oxygenic photosynthesis (Holland 2006; Bekker et al. 2004; Schirrmeister et al. 2013; Schirrmeister, Gugger, and Donoghue 2015; Van Kranendonk et al. 2012). Other dramatic climate shifts later in Earth's history include snowball earth events where the entire Earth was covered in snow; hot houses where no polar ice caps remained and a further oxygenation event; the Neoproterozoic Oxygenation Event (NOE) which raised the levels of atmospheric oxygen to nearly the same as modern day. All of these events might have fundamentally changed the biosphere at the time. More generally the Earth has undergone constant reworking of its crust resulting in shifting continent patterns. A time period where little crust movement occurred from 1.8 to 0.8 Billion years ago is colloquially known as the boring billion where little of

geological note occurred (Brasier and Lindsay 1998). Timescales for the tree of life allow us to consider events such as the GOE, the boring billion and snowball earth events in comparison to the evolution of life.

Despite the record procured from fossils, it is hard to study life's timeline in its entirety. At the very earliest stages of life's evolution the rock record is extremely poor. In fact, there are only a handful of sites across the world that record these time periods and, although this record improves over time, it is still highly imperfect. This means that the oldest fossil that can be confidently assigned to a group is unlikely to represent the age of that group and it would be inappropriate to use it in such a context (Reisz and Muller 2004; Benton and Donoghue 2007). In tandem with this, especially when we are looking at the very earliest splits in the tree of life, it is hard to assign fossils to any particular group of organisms because of the lack of distinct morphological features. The problem of assigning taxonomic affinities continues down the tree and complicates what we can and cannot use to infer life's evolutionary timescale.

In a fossil context we cannot view events such as the mitochondrial endosymbiosis and other key biological changes such as the formation of the chloroplast. Events such as endosymbiosis have come to define a whole scion of the tree of life, Eukaryota, and thus their timing is of huge importance in understanding the evolutionary pathway life has taken. For example, when looking at eukaryotes, how long did the transition from FECA to LECA take and in what order were the cells defining features gained (Embley and Martin 2006; Koonin 2010; Martijn et al. 2018)? In that time period did the mitochondrial endosymbiosis event occur early on in the formation of the cell (Martin et al. 2017) or did it only occur once all the other cell machinery was in place (Pittis and Gabaldón 2016). The change from an archaebacterial cell to a eukaryotic one is profound, and the last eukaryotic common ancestor would have possessed not only mitochondria but also a complex set of cell systems including a nucleus and cytoskeleton attributes (Koumandou et al. 2013).

Once, we have a timescale we can use the results to study the rate of evolution. As each branch on the tree is a combination of rate and time once we have one, we can look at the branch lengths to try and discern something about the other. This theory can be applied to shifts in morphology which often look like they are 'explosive' in the fossil record, for example, the Cambrian explosion. This event could be something which is a product of elevated rates of both morphological and molecular evolution, or there could be a slower rate of molecular change speaking to a complex previous history. Rate information can be used to help analyse the decoupling between the morphological and molecular records (Lee, Soubrier, and Edgecombe 2013). Certainly, molecular clocks on the whole, as well as a now growing body of fossil evidence, place the evolution of metazoans back into the Precambrian (Cunningham et al. 2017; dos Reis, Donoghue, and Yang 2015; Dohrmann and Wörheide 2017).

Due to the reasons outlined above the production of a probabilistic timescale for the tree of life is hugely beneficial to further the understanding of the evolution of life and its relationship to the changing climates and habitats during the evolution of the Earth. Although the fossil record provides a useful insight into the evolution of organisms it is an imperfect record. This is where the probabilistic nature of the molecular clock steps in and allows us to utilise not just the available fossil data but also the wealth of genetic data available from modern organisms. The previously mentioned problems with the fossil record are most acute in the oldest rocks on earth. This patchy record means the most ancient divergences of life, both prokaryotic and eukaryotic, are ideal places to utilise the strength of the combined fossil and genetic data approach of the molecular clock. In this chapter I firstly give a brief overview of the fossil record, before moving on to describe the structure of the tree of life, and finally how to implement molecular clock methods in order to produce a timescale for the tree of life's deep nodes.

1.3 The fossil record of life

The history of life on earth is documented, at least in part, by the fossil record. However, using this record to establish when life might have arisen is tricky (Javaux 2019). The first potential traces of life

appear not long after the formation of Earth and can be found in the very oldest rocks from ~3.8 billion years ago in the Istaq Gneiss in Greenland. These records include putative microfossils (Pflug and Jaeschke-Boyer 1979) (Fig. 1.1 A), stromatolites (Nutman et al. 2016) (Fig. 1.1 C), carbon isotopic signatures (Rosing 1999) and graphite inclusions (Mojzsis et al. 1996; Schidlowski 1988). Records of possible microfossils of a similar age are also found in Nuvvuagittuq, Canada (Dodd et al. 2017). However, these records are all inherently controversial because their biological affinity is difficult to substantiate, and, in some cases, it is equally if not more probable that they were produced geologically. If we want to provide a solid timescale that is not subject to constant rewriting, then we need to be conservative in our approach to the assessment of fossil material. This means that although these records have some merit, they cannot be conclusively assigned a biological affinity. The microfossils in both cases have not been rigorously assessed and the stromatolites might have a biological origin, but similar structures have been produced in the lab (McLoughlin, Wilson, and Brasier 2008; Grotzinger and Rothman 1996). Additionally, they have no associated microfossils to help confirm their biogenicity. The rocks at this site have also undergone a degree of metamorphism meaning it is harder to be certain of the validity of the structures we see within them. Similar problems of biological authenticity plague the isotopic signatures which can also be produced by abiogenic sources, such as the Fischer-Tropsch type reactions (Lollar et al. 2002; Horita and Berndt 1999) where very negative carbon isotope signatures can be produced without biological involvement, thus cannot be used as conclusive proof for life. The next putative evidence for life appears in the Pilbara craton, Australia (~3.4 Ga) in the Dresser and Strelley Pool Formations. These sites have better candidates for biologically produced structures. In particular the Strelley Pool Formation, Pilbara, has chains of cells (Sugitani, Mimura, Takeuchi, Yamaguchi, et al. 2015) (Fig. 1.1 C, D) associated with rocks that also possess more rigorously assessed stromatolites (Wacey 2010) as well as other microfossils (Wacey et al. 2011; Sugitani et al. 2013) (Fig. 1.1 E). These combined records help to establish the oldest fossil evidence for life.

Subsequent to the establishment of life 3 billion years ago, the life forms on our planet and thus the fossil record were dominated by prokaryotes, Eubacteria and Archaebacteria, generally single celled, though possessing a huge range of metabolisms allowing them to conquer numerous environments

(Nowak and Knoll 2017). Fossils from these early time periods are difficult to assign to any extant clades owing to the lack of morphological features available for comparison. One of the most significant changes in environment during this time is the Great Oxidation Event (GOE) (Holland 2006; Bekker et al. 2004) a sustained increase in atmospheric oxygen levels that happened around 2.4 – 2.1 Ga (Fig. 1.2). As Cyanobacteria are the only lineage primarily capable of oxygenic photosynthesis, they have typically been associated with the GOE (Schirrmeister et al. 2013; Schirrmeister, Gugger, and Donoghue 2015; Bekker et al. 2004; Van Kranendonk et al. 2012). However, there is some debate as to where in the crown or stem lineage this characteristic arose and when Cyanobacteria actually evolved (Schirrmeister et al. 2013; Schirrmeister, Gugger, and Donoghue 2015; Betts et al. 2018; Shih and Matzke 2013).

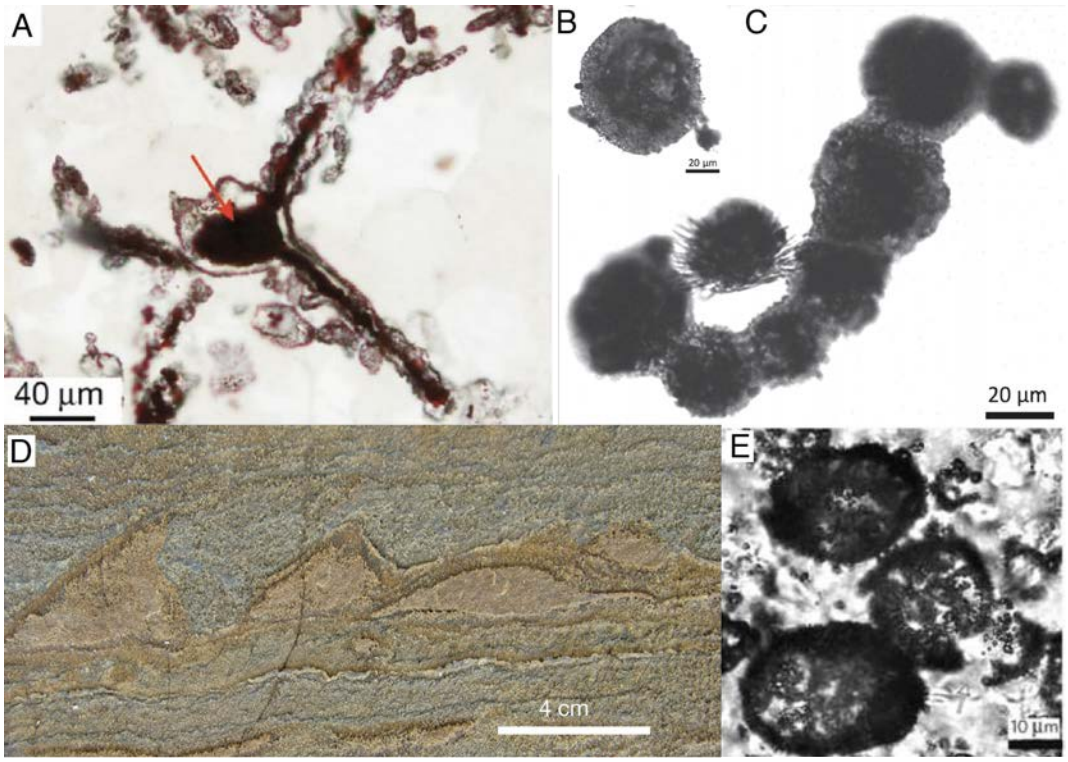


Figure 1.1. Examples of ancient putative cells. A) Haematite filaments thought to be similar to modern microfossils, Nuvvuagittuq, Canada (Dodd et al. 2017) the red arrow indicates the terminal of the filaments, B) and C) examples of cells from the Pilbara craton, Strelley Pool Formation (Sugitani, Mimura, Takeuchi, Lepot, et al. 2015) , D) putative stromatolites from the 3,800 Ma Isua Supracrustal Belt (Nutman et al. 2016) and E) possible sulphur metabolising microfossils from the Strelley Pool Formation (Wacey et al. 2011).

The GOE occurred in tandem with an increasing number of potential cyanobacterial fossils (Altermann and Schopf 1995; Klein, Beukes, and Schopf 1987; Knoll, Strother, and Rossi 1988; Hofmann 1976; Amard and Bertrand-Sarfati 1997), although these cannot be conclusively linked to this clade. There are also earlier 'whiffs' of oxygen (Satkoski et al. 2015; Anbar et al. 2007; Crowe et al. 2013; Kendall et al. 2010; Czaja et al. 2012; Planavsky et al. 2014; Riding, Fralick, and Liang 2014) which speak to a lag between the possible evolution of Cyanobacteria and their effect upon the atmosphere.

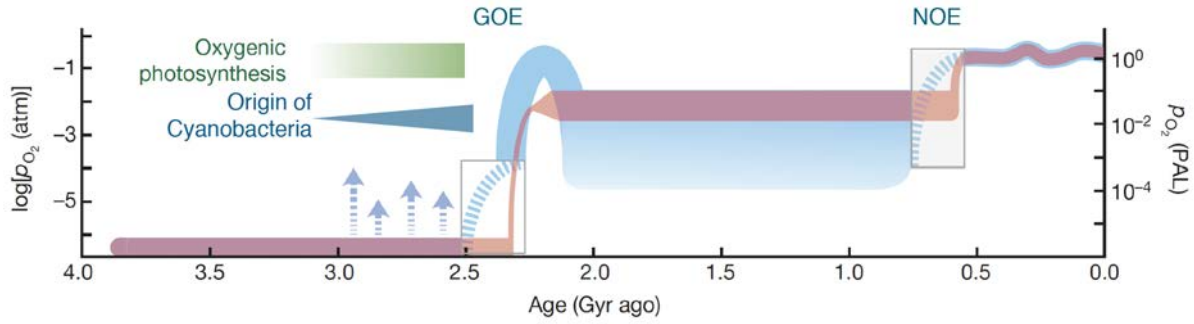


Figure 1.2. A timeline showing the increase in oxygen throughout Earth's history (red curve) (modified from (Lyons, Reinhard, and Planavsky 2014)). The blue curve indicates the model of pO2 = atmospheric partial pressure of O2. The right-hand axis shows the level of pO2 relative to modern day (partial pressure of O_2; PAL) and the left-hand axis shows log PO_2. GOE = Great oxidation event and NOE = Neoproterozoic oxygenation event.

Just as when we are looking at records of very early life, we cannot be certain of some of the taxonomic affinities of the earliest potential records of eukaryotes. Perhaps the oldest fossil claimed to be of eukaryotic ancestry is *Grypania spiralis* from ~2.1 Ga (Han and Runnegar 1992) (Fig. 1.3 A). However, this has also been compared to Cyanobacteria (Sharma and Shukla 2009) making it difficult to establish *Grypania's* direct affinities. The first definite eukaryote fossils begin to appear in the Proterozoic around 1.8 – 1.6 Ga. These fossils fall into the category of acritarch, a catch-all name for single celled organisms which are most probably of eukaryotic ancestry, but which, beyond that, have uncertain phylogenetic affinities. It is likely that the oldest acritarch fossils which belong to eukaryotes are from the Changzhougou System (Peng, Bao, and Yuan 2009; Lamb et al. 2009; Knoll and Nowak 2017). The cell wall ultrastructure (Fig. 1.3 B) coupled with their large size (not an indicative feature on its own) mean that they are most likely eukaryotes. Other acritarch forms appear in similarly dated formations

from China (Ruyang group (Leiming et al. 2005; Yin and Yuan 2003)) and Australia (Roper group (Javaux, Knoll, and Walter 2001; Javaux, Knoll, and Walter 2004; Yin 1997)). These are usually simple forms, sometimes with some cell wall ornamentation and processes distinguishing species such as *Tappania plana* (Fig. 1.3 C) and *Valeria lophostriata* (Fig. 1.3 D). Despite their features none of these fossils can be allied with any extant eukaryotic group and thus they can at best be used to give an indication of when the total group had evolved.

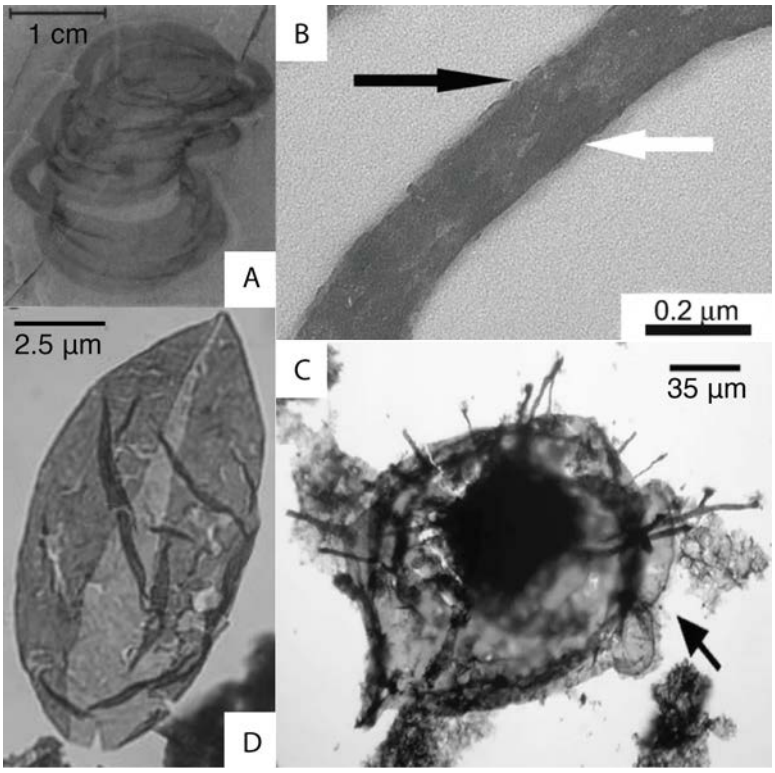


Figure 1.3. Early possible and probable eukaryote fossils, A) *Grypania spiralis (Han and Runnegar 1992)*, B) Acritarch cell wall from Changzhougou, arrows show two sides of the trilaminar wall (Peng, Bao, and Yuan 2009), C) *Tappania plana* (Javaux, Knoll, and Walter 2001) and D) *Valeria lophostriata* (Javaux, Knoll, and Walter 2004).

The first recognisable crown group eukaryote is often thought to be *Bangiomorpha pubescens* dating to a minimum age of 1.03 Ga (Butterfield 2000). However, this has recently been challenged by older potential crown group rhodophytes described from the ~1.6 Ga Chitrakook Formation (Bengtson, Sallstedt, et al. 2017). Other ancient crown eukaryote fossils include possible fungi that are slightly younger than *Bangiomorpha* ~ 1 Ga (Loron et al. 2019) and other algal groups (Xiao et al. 2004;

Butterfield 2004; Butterfield, Knoll, and Swett 1994). Collectively, these fossils push the origin of crown eukaryotes at the very least back into the Mesoproterozoic (> 1000 Ma), and signify that both the chloroplastic, and mitochondrial endosymbiosis events had already occurred. Later we see testate eukaryotic forms, known as vase-shaped microfossils (Bosak, Macdonald, et al. 2011; Porter, Meisterfeld, and Knoll 2003) which could be either ameobozoans or part of the SAR grouping (stramenopiles, alveolates and rhizarians). At the very end of the Proterozoic the enigmatic Ediacaran fauna and the first signs of metazoan life are recorded in the rocks (Fedonkin et al. 2007). There is a body of work suggesting that some fossil groups from the Ediacaran biota belong within Metazoa (Hoekzema et al. 2017; Dunn, Liu, and Donoghue 2018). However, it is not until the Phanerozoic in the Cambrian period that we see a full flourishing of metazoans. This occurs in an apparent explosion both in terms of the number of fossils and the diversity of forms with all the major metazoan clades appearing in the fossil record. Although the idea of an explosion is being increasingly challenged and it seems more likely that many metazoan groups were present prior to this time (Cunningham et al., 2017, Wood et al., 2019). The Phanerozoic sees important additions to the fossil record for example flowering plants (Morris et al. 2018), tetrapods (Benton et al. 2013) and insects (Labandeira 2018).

Although this wealth of fossil data does allow us to roughly track the evolution of life through time our efforts to establish a timeline are hampered by the patchy and poor rock record. Not every environment is as likely to be fossilised and in each environment the organisms have different fossilisation potentials. This is coupled with problems associated with the fossilisation potential of the organisms themselves. Eukaryotes that possess hard parts, such as arthropods, molluscs and tetrapods are more likely to fossilise because of the tough parts of their anatomy. By contrast it is much harder to find specimens of exclusively soft bodies organisms. Additionally, there are layers of rock we cannot access either because they are not present at the surface, or because they have already been reworked by the Earths rock recycling processes. All of these problems mean that while useful the fossil record does not provide a complete overview of life's evolutionary trajectory.

1.4 The molecular clock

The molecular clock is a technique that was first introduced in the 1960s by Zuckerkandl and Pauling (1962, 1965). They proposed that the number of substitutions fixed between homologous amino acid sequences in mammals was roughly equivalent to the time since the species had shared a common ancestor. This, coupled with information from the fossil record, allows the dating of species divergences for which no fossil information is available. The rate of substitution was thought of as stochastic where mutations occur at random and are thought of as the 'tick' rate of the clock. Initial versions of the molecular clock were 'strict', they assumed that evolutionary rate was constant through time and they applied a maximum likelihood approach with the fossils incorporated as point estimates. However, it was quickly realised that the strict clock did not hold true for most data sets.

Newer, more complex molecular clock models rely on a Bayesian framework and Markov Chain Monte Carlo (MCMC) methods. Bayes theorem was first applied to molecular clocks around the turn of the century (Thorne, Kishino, and Painter 1998) and Bayesian estimation of divergence times allows the explicit incorporation of a number of uncertainties into the analysis. These can be constrained using a prior reflecting what we already know about the subject. This prior information then interacts with the data to produce the posterior probability distribution (dos Reis, Donoghue, and Yang 2015). We now know that the strict clock does not hold true in most circumstances and Bayesian methods have been developed to relax the assumption of a constant substitution rate amongst species. Relaxed clock models fall into two main categories, one where the rate of evolution on each branch can be different (Drummond et al. 2006; Rannala and Yang 2007), the other where each branch has a rate related to its parent branch (Kishino, Thorne, and Bruno 2001; Thorne, Kishino, and Painter 1998; Thorne and Kishino 2002; Lepage et al. 2006) an idea originally floated in a 1991 work by Gillespie (Gillespie 1991) who first suggested that the molecular clock should logically be best seen as an autocorrelated process. The latter means that you end up with rates of evolution that are more similar within lineages. These types of clock models are known as uncorrelated and autocorrelated relaxed clocks respectively and their specification sets a prior on the rates. Some authors have proposed uncorrelated models to be

superior (Drummond et al. 2006; Linder, Britton, and Sennblad 2011) and others have argued in favour of autocorrelation (Lepage et al. 2007). The choice of which kind of relaxed clock model to use is most likely dependent on the individual datasets being analysed.

The molecular clock can be influenced by a number of factors such as convergent evolution, fast evolving genes, generation time and strength of selection as well as duplications, losses or horizontal gene transfers which can be important when trying to infer the underlying tree topology (Bromham 2019). All of these factors can mean that the 'tick' rate for those species is distorted. Therefore, when we use molecular clocks, we have to be acutely aware of the assumptions we are making and thus the potential sources of error we are incorporating into our dataset (Bromham 2019).

1.5 The fundamental structure of the tree of life

In order to exploit molecular clocks, we can apply them to a previously constructed phylogenetic tree. This means a prior knowledge of the relationships between the included species must be known, either through the previous work of other authors, or ad hoc production of a phylogenetic tree using appropriate methods. Here too there are a number of possible production methods ranging from parsimony to Bayesian estimation. Currently the most favoured are maximum likelihood and Bayesian. The latter once again allows for the incorporation of prior information, and the implementation of more complex models, and produces posterior probabilities.

It is generally accepted that life on Earth shares a last universal common ancestor (LUCA) and that there are two major kinds of life, organisms with a nucleus, Eukaryota, and organisms without a nucleus, Prokaryota, which is further broken down into Archaebacteria and Eubacteria. LUCA is a somewhat enigmatic organism proposed to have formed in a variety of environments ranging from a prebiotic soup (Haldane 1929) to fiery (Corliss 1981; Baross and Hoffman 1985) or warm alkaline hydrothermal vents (Martin and Russell 2006; Russell, Hall, and Martin 2010). It has been suggested to have a selection of, at the very least, 355 protein families (Weiss et al. 2016) and a potentially anaerobic

metabolism involving H_2-dependency and the Wood-Ljungdhal pathway (Weiss et al. 2016; Weiss et al. 2018). Exactly how LUCA's descendant groups of organisms are related is an ongoing debate in molecular phylogenetics and it has changed our view on how prokaryotes and eukaryotes related to each other. Developments over time in sequencing techniques and ways to process the sequences have led to 4 main schools of thought about the topology of the tree of life summarised in (McInerney, O'Connell, and Pisani 2014). These 4 ideas are outlined below; the three domains hypothesis, the Eocyte hypothesis, the ring of life hypothesis and the eukaryote early hypothesis (Fig. 1.4).

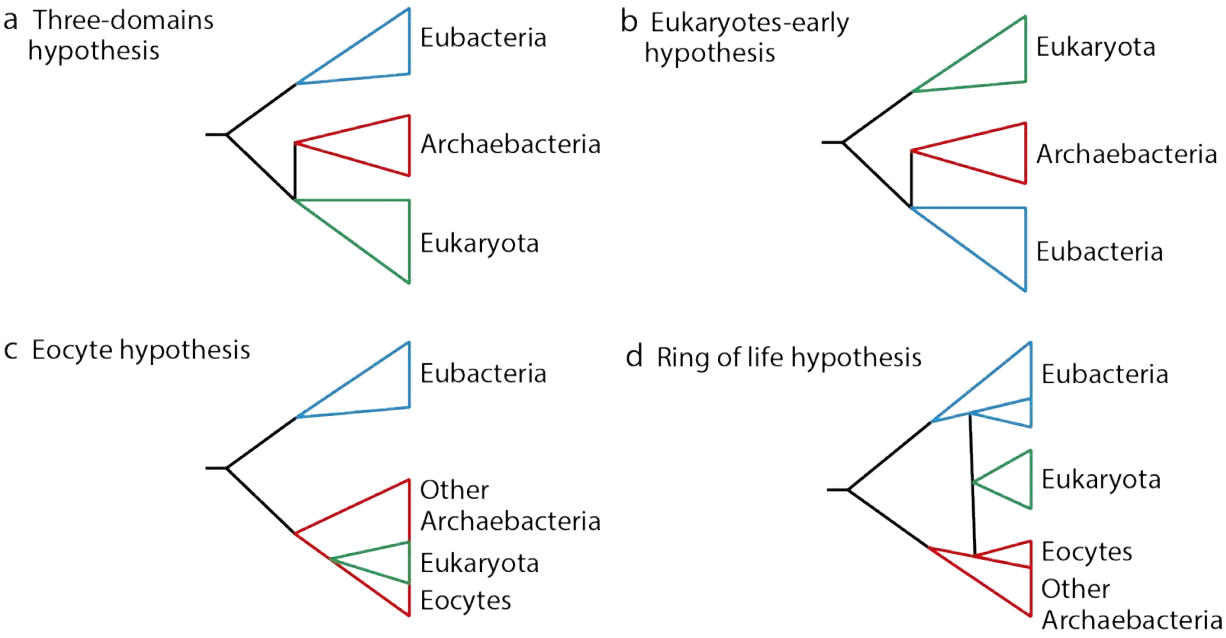


Figure 1.4. The different views on the structure of the tree of life (Modified from (McInerney, O'Connell, and Pisani 2014)). a) The three-domains hypothesis, b) the eukaryote early hypothesis where the prokaryotes arise from a eukaryotic ancestor, c) the Eocyte hypothesis where the eukaryotes arise as a lineage within the archaebacteria and d) the ring of life hypothesis where the ancestral eukaryote cell formed as a merger between an archaebacterial host and a eubacterial endosymbiont.

Before the advent of sequencing technology little was known about the evolution of life especially the prokaryotes because their simple morphologies did not allow for detailed comparisons. There existed an idea of a prokaryote-eukaryote split, without the acknowledgement of Eubacteria and Archaebacteria as distinct lineages. Archaebacteria were not considered a separate group until 1977 (Woese and Fox

1997). They were discovered when an analysis of ribosomal RNA sequences found that what was previously known as methanogenic Eubacteria, were in fact very distinct from other eubacterial lineages (Woese and Fox 1977). Despite sharing simple morphologies with no nucleus Eubacteria and Archaebacteria are fundamentally different, something which molecular studies have helped to elucidate. This was achieved partially through the use of gene duplications which occurred before LUCA, each duplicate could be used to root the other, and search for the root of the tree of life (Gogarten et al. 1989). This analysis and others (Pace, Olsen, and Woese 1986) found that Eukaryota and Archaebacteria were more closely related to each other than either were to Eubacteria. The three domains idea was formalised by Woese and colleagues who envisaged a system where each domain was monophyletic (Woese, Kandler, and Wheelis 1990).

Later it was noted that the ribosomal structure of eukaryotes and some archaebacterial lineages was similar, specifically the Crenarchaeota, and thus the idea of eukaryotes arising within Archaebacteria was formed (Lake et al. 1984). This is known as the Eocyte hypothesis. Studies involving elongation factors (Baldauf, Palmer, and Doolittle 1996; Rivera and Lake 1992) strengthened the idea that eukaryotes could be related to Crenarchaeota, as did supertree methods with genomic scale support (Pisani, Cotton, and McInerney 2007) and ribosomal RNAs with gene concatenation methods (Cox et al. 2008). Updated phylogenetic techniques and improved sampling of archaeal genomes subsequently found that eukaryotes might lie next to or within the TACK Archaebacteria (the superphylum traditionally containing Thaumarchaeota, Aigarchaeota, Crenarchaeota and Korarchaeota) (Guy and Ettema 2011; Williams et al. 2012; Williams et al. 2013) supported by cell features involved in cytokinesis and membrane remodeling (Hurley and Hanson 2010), and cell shape determination (Ettema, Lindås, and Bernander 2011). The Eocyte hypothesis has additionally been strengthened by the relatively recent discovery of the Asgardarchaeota (Spang et al. 2015; Zaremba-Niedzwiedzka et al. 2017), a group of Archaebacteria only known from metagenomic sequencing, but which possess genes originally thought to be exclusive to eukaryotes, such as signature proteins and cytoskeleton components. When included in phylogenies eukaryotes align with this newly discovered group, either as sister to, or within Asgardarchaeota. This finding has been challenged owing to the nature of some

of the genes used in the sample and the metagenomic sampling of the original specimens (Da Cunha et al. 2017). However, these criticisms have been robustly rebuffed (Spang et al. 2018), and it now seems certain that the ancestral eukaryote was derived from an archaeal cell. This is additionally supported by a phylogeny produced from a huge sample of extant lineages which even finds that the Eubacteria might have two distinct groupings, all previously known Eubacteria, and the newly discovered candidate phyla radiation (Hug et al. 2016).

The ring of life describes the eukaryote cell as an integration (a “fusion” in the terminology of Rivera and Lake (Rivera and Lake 2004) of an archaeal and bacterial cell through a flow of genetic material involving both the archaebacterial host and the mitochondrial endosymbiont (Rivera and Lake 2004; McInerney, Pisani, and O'Connell 2015). Almost all eukaryotes possess mitochondria although in some species the organelle has been highly modified and thus bears little resemblance to its original form or in extreme cases has been lost (Karnkowska et al. 2016). The mitochondrion is generally thought to have originated via the engulfing of an alphaproteobacterial cell by an archaeal one. Numerous studies have related the core of the mitochondrial genome to the aforementioned bacterial lineage (Bonen et al. 1977; Schwartz and Dayhoff 1978; Yang et al. 1985; Wang and Wu 2015; Rodríguez-Ezpeleta and Embley 2012), but mitochondria also possess genes from other Eubacteria as well as eukaryotes and genes of their own (Roger, Muñoz-Gómez, and Kamikawa 2017; Gray 2015). Which alphaproteobacterial lineage gave rise to the mitochondrion is debated. The organelle been linked to the Rickettsiales (Wang and Wu 2015, 2014; Fitzpatrick, Creevey, and McInerney 2005), but this lineage has small genomes and is often parasitic which might be causing a long branch attraction effect. When accounted for this seems to suggest that actually the ancestral mitochondrion is derived from a different alphaproteobacterial lineage (Abhishek et al. 2011; Thiergart et al. 2012) or even as an independent branch in the group (Rodríguez-Ezpeleta and Embley 2012; Martijn et al. 2018; Esser et al. 2004).

1.6 Dating for the tree of life

1.6.1 Node calibration

Chapter 2. In this chapter I review the fossil material available in order to calibrate the tree of life and then find divergence dates for it using node calibration. To date the tree of life using a molecular clock there are a few methods available, one of the most well established is node calibration. In this method fossil calibrations are incorporated into the molecular clock model via the prior on times. The fossils provide a minimum calibration for a group when a given fossil is known to be the oldest representative of that particular lineage. They help us to anchor the tree in real time, without them we can gain insight into the relative ages of the chosen species (Loader et al. 2007) but not the absolute ages with which we can make inferences about life and the geological record. The calibration must be carefully constructed using a minimum date that reflects the geological age of the fossil. A maximum can also be applied which in most cases is formed in a conservative manner by the absence of evidence of any fossil of that particular group. The maximum bounds of calibrations can be implemented in a 'soft' manner (Yang and Rannala 2006), meaning that the analysis can break the bounds of this calibration if the data overrides it, usually by a given percentage e.g. 2.5%. This is the case because maximums are generally based only on a lack of evidence. The best way to construct a node calibration was laid out in 2011 (Parham et al. 2011) in which they suggest that calibrations should be constructed in two parts, the justification of the phylogenetic position of a fossil/record, and the justification of the age of that fossil/record. It is important to correctly construct calibrations because they can have major effects on the divergence times produced (Warnock et al. 2015). For the calibration to be considered robust it requires; an up to date phylogenetic appraisal, a knowledge of what rock layer that fossil or isotope record sits in, and a radioisotopic date for near the fossil or a date for a comparable rock layer as agreed by the international geological timescale. For example, if we want to set a minimum constraint on the age of life then we can use the oldest fossil evidence (Fig. 1.5A and 1.5B) which comes from the Strelley Pool Formation (Sugitani et al. 2013; Sugitani, Mimura, Takeuchi, Yamaguchi, et al. 2015; Sugitani, Mimura, Takeuchi, Lepot, et al. 2015). This Formation underlies the Euro Basalt which is associated with a tuff dated to 3350 Ma ± 3 Myr (Nelson 2005). Hence, the minimum age for any fossil at this

formation is 3347 Ma. We can then set an upper bound by looking for the earliest date life could possibly have existed. In this case it is the moon forming impact, something which would have fundamentally reworked the planet and which no life could have survived, which provides a maximum date of 4520 Ma (Fig. 1.5C). This calibration as well as others useful for calibrating the tree of life were formally laid out in Betts et al., 2018.

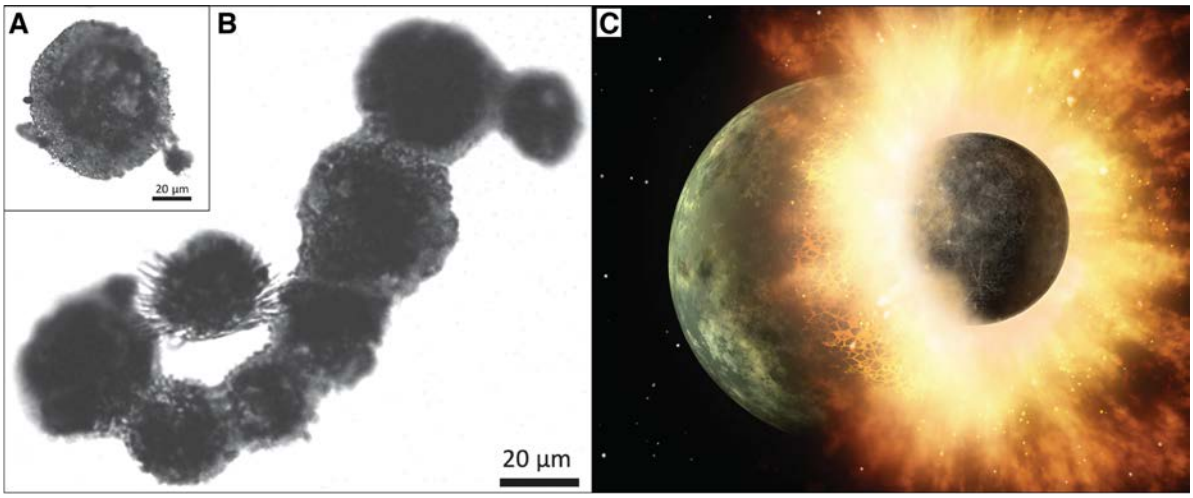


Figure 1.5. The constraints on the age of life. A,B) Fossils from the Strelley Pool Formation (Sugitani, Mimura, Takeuchi, Lepot, et al. 2015). A) Smooth walled, partially centrally hollow fossil with granular flange and B) Chain composed of 7 individual cells. C) An image of the moon forming impact, a massive collision between the proto-earth and the planetary body known as Theia.

Another very important aspect of node calibration is modelling how the calibrations are implemented onto their calibrated node. This means testing how well the calibrations minimum bound, the age of the oldest representative fossil, corresponds to the actual node age. The calibration density distribution can be manipulated so that the fossil is a good, moderate or poor estimate of node age, or, it can reflect that we have no distinct idea of where the node lies, thus employing a uniform distribution. For example, if the oldest fossil evidence for life appears at roughly 3350 Ma then we can say whether we think that is likely to be close to when LUCA was present, or we can specify that life likely evolved earlier, closer to the upper bound, the moon forming impact. In practice there is often very little information to be able to choose one of these scenarios conclusively and thus it could be better to integrate over the uncertainty associated with all approaches (Betts et al., 2018).

1.6.2 Cross calibration and its application in dating nodes close to the root of a tree

Chapter 3. In this chapter I utilise gene duplications in a node calibration framework to better estimate dates for the last universal common ancestor. In a traditional node calibration analyses the node we have the least information for is always the root node. In the case of the tree of life this node is LUCA. In order to have more information about such a node we need an outgroup and one way we can find this is using gene trees which have a duplication prior to the root of the tree of life. This means that each paralog can root the other. Duplicated gene datasets were originally used to find topologies for the tree of life where the main duplication is right at the root of the tree prior to the last universal common ancestor. These studies were spearheaded by Gogarten and colleagues by an analysis using vacuolar H^+-ATPases (Gogarten et al. 1989). This was then replicated by Iwabe et al., 1989 who used elongation factors and ATPases (Iwabe et al. 1989) and again later using both elongation factors (Baldauf, Palmer, and Doolittle 1996). All corroborated the placement of eukaryotes next to or within Archaeabacteria and provided a useful mechanism to root for the tree of life. Here, we exploit this methodology in order to find divergence times for the tree of life, specifically for nodes close to the root which otherwise we would have less information to resolve.

In duplicated trees there are two kinds of node, duplications and speciations. Speciation nodes on either side of the duplications are from the same event and so can be calibrated using the same prior information just like in a normal node calibration analysis. Two methods can be implemented, either cross-calibration where the same prior is used but each node can have its own posterior distribution (Clark and Donoghue 2017) or, in the case of cross-bracing, we can force the nodes to have the same output meaning that information from both sides of the duplication is used to inform the age of the node (Fig. 1.6). This technique was first used when looking at endosymbiosis events (Shih and Matzke 2013). There are only a handful genes that are both present before the last universal common ancestor and which have a duplication event prior to its evolution. These genes are the vacuolar H^+ATPases, elongation factors EF-Tu/1 and EF-G/2, the Histidine biosynthesis subunits A and F, aspartate and ornithine carbamoyltransferases, signal recognition proteins, tyrptophanyl-tRNA and tyrosyl-tRNA synthetases and, finally, the Valyl-, Methionyl-, Isoleucyl- and Leucyl- tRNA synthetases. These genes

can be used to date the tree of life and more specifically LUCA, both as individual genes and also in a combined analysis.

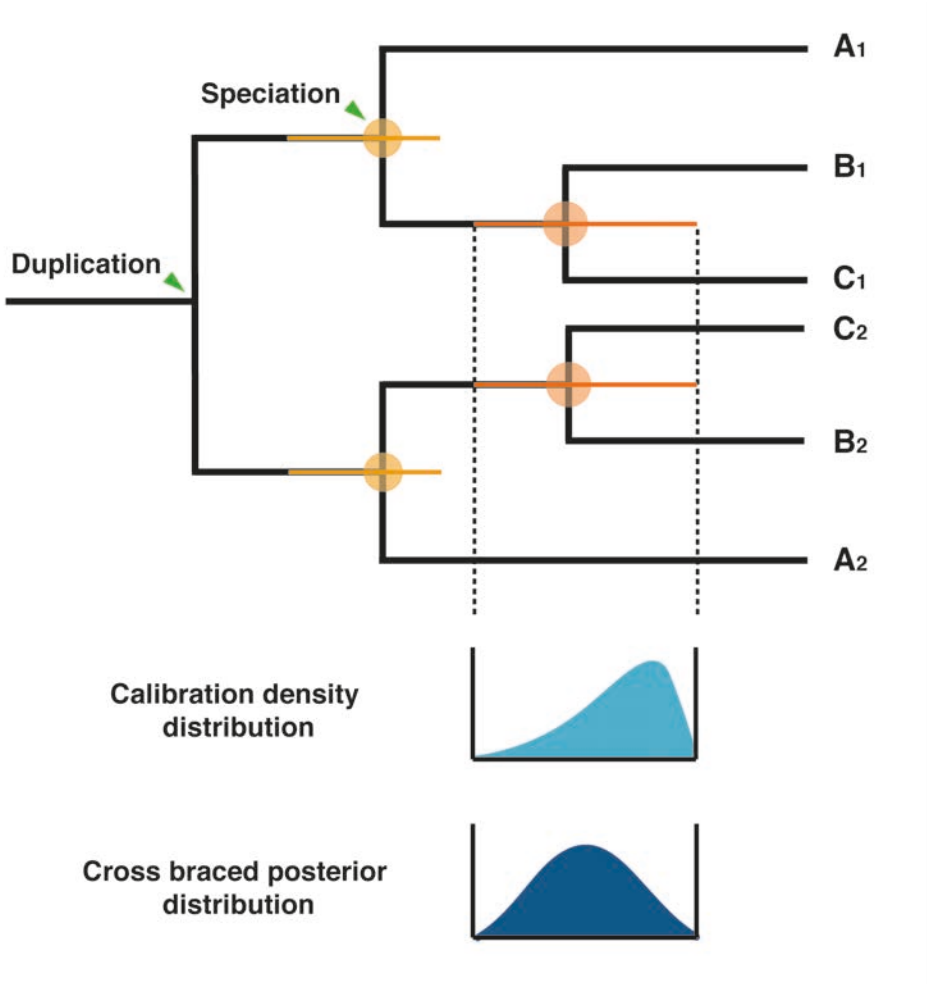


Figure 1.6. Cross-bracing of a duplicated gene tree. In this case there is one duplication event at the root of the tree and then two subsequent speciation events to end up with 6 tips, A1 and A2 are paralogs and A1 and B1 are orthologs. For each speciation event the same prior calibration density distribution can be applied (light blue curve), and the same posterior density distribution is produced (dark blue curve).

1.6.3 The fossilised birth-death process

Chapter 4. In this chapter I introduce the Fossilised Birth Death process and describe its use for dating ancient divergences with the study group of eukaryotes. When we use node calibration for divergence time estimation, we have to be absolutely certain of the phylogenetic placement of each fossil within the given topology, in order to produce a robust and accurate prior (Parham et al. 2011; Warnock et al. 2015). This is possible for groups with good fossil records and easily recognisable diagnostic features.

However, many fossils cannot be assigned with confidence to any one node, or in fact any particular group, and, even if there is more than one fossil record for a given node, within a node calibration framework only the oldest can be used. This means that we lose a great of information from the fossil record, especially in older time periods where fewer fossils can be confidently assigned to a clade for example. The problems with the phylogenetic uncertainty of fossils and this lack of integration of such uncertainty into node calibrated analyses mean that some people prefer to use alternative methods.

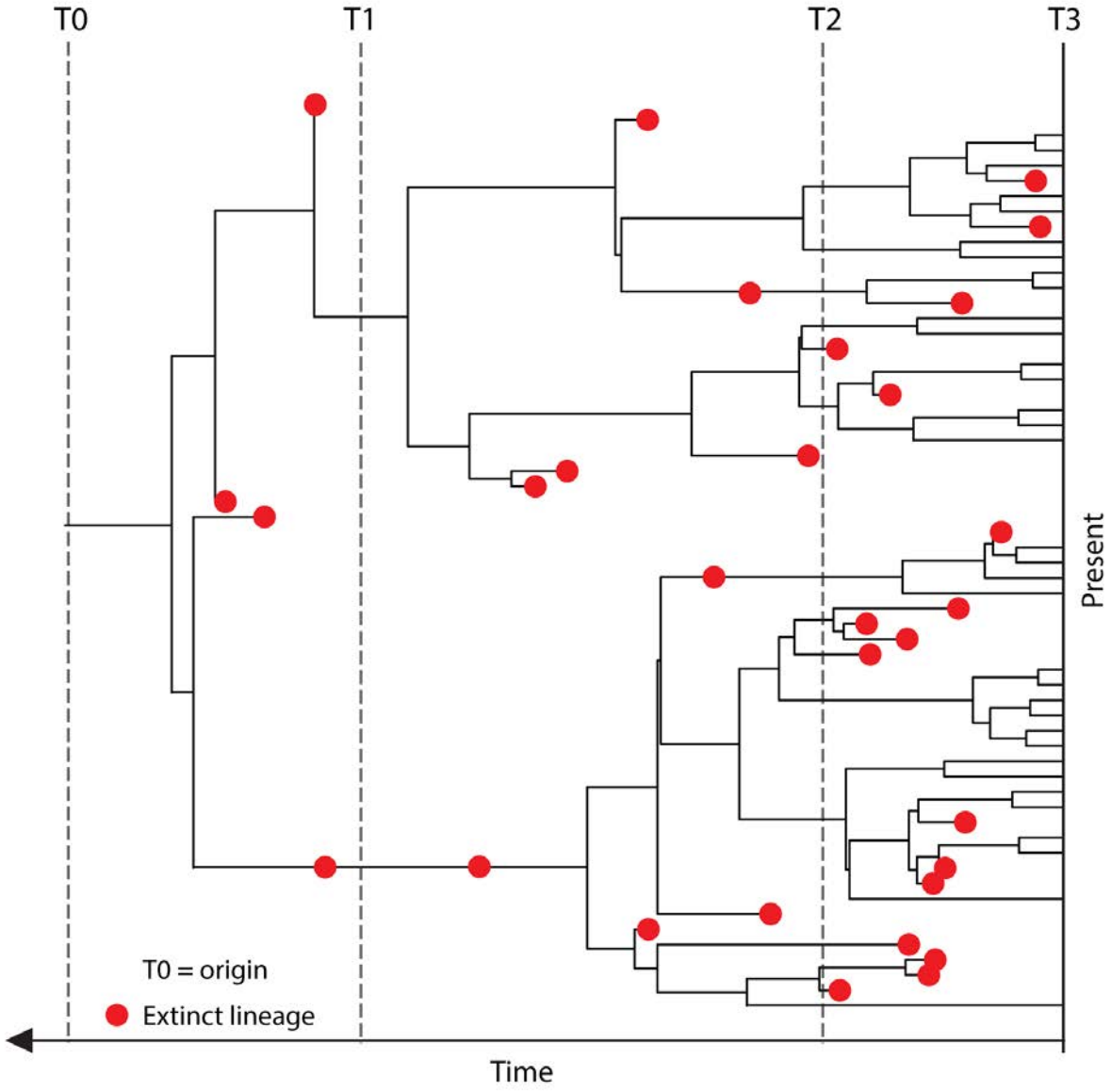


Figure 1.7. An illustration showing key parts of the fossilised-birth-death skyline model from the origin of the process, T0 to the present T3. This model allows for the specification and estimation of different birth, death and fossil sampling rates within each of the time bins. Hence, the rates can vary between times, but not between lineages.

One of the newest methods of total-evidence dating incorporates fossils explicitly into the prior on nodes times. Called the fossilised-birth-death (FBD) process (Stadler 2010; Heath, Huelsenbeck, and Stadler 2014) it is a relatively recent expansion to the more widely used birth-death model which provides a measure of the rate at which species arise and become extinct. The FBD accounts for the process of fossilisation and how likely this is. This means that if computationally possible many more fossils can be included into the analysis giving a better overview of the complete evolution of the lineage. It can make use of both molecular and morphological data in order to place the fossils. If morphological data are used, then the analysis is resolved. However, if the latter is unavailable then the analysis marginalises over all the fossil attachment points (Heath, Huelsenbeck, and Stadler 2014) and is known as the unresolved FBD. This implementation of the FBD process requires that the fossils are constrained within the extant phylogeny, though stem and crown affinities can both be specified, making this approach less strict than that of node calibration. Extensions to this model also allow the incorporation of sampled ancestors (Gavryushkina et al. 2014) and time slicing of the analysis (Stadler et al. 2013) (Fig. 1.7). The time slicing of analyses can be useful if, for example, we know that the number of fossils is much smaller in the earliest part of the tree.

Chapter 2

Integrated genomic and fossil evidence illuminates life's early evolution and eukaryote origins

Author contributions: This chapter is published in *Nature Ecology and Evolution*.

Betts H.C., et al., 2018. Integrated genomic and fossil evidence illuminates life's early evolution and eukaryote origin. *Nature Ecology & Evolution.* 2, 1556–1562.

The ideas for this chapter were developed by D. Pisani, P.C.J. Donoghue, T.A. Williams, H.C. Betts, M.N. Puttick and J.W. Clark. The dataset was collated by H.C. Betts and molecular analyses were carried out by H.C. Betts, M.N. Puttick and J.W. Clark. Calibrations were researched, developed and written by H.C.B with advice from P.C.J.D. H.C.B wrote the paper with comments and suggestions from the other authors. H.C.B is the lead author of this work and contributed to 90% of the work presented here.

2.1 Introduction

Establishing a unified timescale for the early evolution of Earth and life is challenging and mired in controversy because of the paucity of fossil evidence, the difficulty of interpreting such evidence and dispute over the deepest branching relationships in the tree of life. Surprisingly, it remains perhaps the only episode in the history of life where literal interpretations of the fossil record hold sway, revised with every new discovery and reinterpretation and attempts to investigate the emergence of life and its subsequent evolution have traditionally focused on the fossil record. However, this record, especially when looking at the earliest scions of life, is minimal and interpretation is made harder due to difficulties substantiating relationships within the earliest branching lineages of the tree of life (dos Reis, Donoghue, and Yang 2015; Wacey 2009; Javaux 2019). Despite its problematic nature, the fossil record remains the main source of information for the time- line of life's evolution. We attempt to shed light on this early period by presenting a molecular timescale based on the ever-growing collection of genetic data, and explicitly incorporating uncertainty associated with fossil sampling, ages and interpretations (dos Reis, Donoghue, and Yang 2015; Parham et al. 2011; Warnock et al. 2015; Inoue, Donoghue, and Yang 2009).

Calibrations are a crucial component of divergence time estimation. Relative divergence times can be inferred using alternative lines of evidence; for example, lateral gene transfers (Davín et al. 2018). However, an absolute timescale for evolutionary history can only be derived when calibrations are included in the analyses (Lozano-Fernandez et al. 2017; Pisani and Liu 2015). We derived a suite of calibrations, following best practice (Parham et al. 2011) for the fundamental clades within the tree of life, drawing on multiple lines of evidence, including physical fossils, biomarkers and isotope geochemistry (Wacey 2009). Two key calibrations, for the last universal common ancestor (LUCA) and the split between Archaebacteria and Eukaryota, constrain the whole tree by setting a maximum on the root, while also informing the timing of divergence of eukaryotes within Archaebacteria (Spang et al. 2015; Williams et al. 2013). Putative records for life extend back to the Eoarchaean, including microfossils (Dodd et al. 2017; Pflug and Jaeschke-Boyer 1979), stromatolites (Nutman et al. 2016)

and isotope data (Mojzsis et al. 1996; Rosing 1999) from the ~3.8 billion years ago (Ga) Isua Greenstone Belt (Greenland). However, these records have been contested (van Zuilen, Lepland, and Arrhenius 2002; Horita and Berndt 1999; Lepland, Arrhenius, and Cornell 2002). Microfossils from the ~3.4 Ga Strelley Pool Formation, Australia, are the oldest conclusive evidence to constrain the age of LUCA (Sugitani, Mimura, Takeuchi, Lepot, et al. 2015). The fossils, many of which are arranged in chains of cells, have been shown, through nanoscale imaging and Raman spectroscopy, to exhibit a complex morphology with a central, usually hollow, lenticular body and a wall that is either smooth or in some cases reticulated; these features are beyond the scope of pseudofossils (Wacey 2009). The Strelley Pool Formation also contains other microfossils (Sugitani et al. 2010; Sugitani et al. 2013; Wacey et al. 2011), in association with both distinct $\partial 13C_{org}$ and $\partial 13C_{inorg}$ (Lepot et al. 2013) and pyrite indicative of sulfur metabolisms (Wacey et al. 2010), along with stromatolites that exhibit biological structure (Wacey 2010). Overall, these data allow us to confidently use the Strelley Pool Biota as the oldest, undisputable, record of life. For a maximum constraint on the age of LUCA, we considered the youngest event on Earth that life could not have survived. Conventionally, this is taken as the end of the episode of late heavy bombardment, but modelling has shown that this would not have been violent enough for planet sterilization (Abramov and Mojzsis 2009). However, the last formative stage of Earth's formation - the Moon-forming impact - melted and sterilized the planet. The oldest fossil remains that can be ascribed to crown Eukaryota are ~1.1 Ga *Bangiomorpha pubescens* (Butterfield 2000; Sánchez-Baracaldo et al. 2017), which can be confidently assigned to the red algal total group (Rhodophyta). Older fossil remains from the >1.561 Ga Chitrakoot Formation have been tentatively interpreted as red algae (Bengtson, Sallstedt, et al. 2017); however, current knowledge of their morphology does not allow for an unequivocal assignment to crown Archaeplastida. The oldest fossil remains that can be ascribed with certainty to total-group Eukaryota are acritarchs from the >1.6191 Ga Changcheng Formation, North China (Lamb et al. 2009), which are discriminated from prokaryotes by their large size (40–250 μm) and complex wall structure, including striations, longitudinal ruptures and a trilaminar organization. However, these structures do not indicate membership of any specific crown eukaryote clade, only allowing us to use these records to minimally constrain the timing of divergence between the Eukaryota and their archaebacterial sister lineage, Asgardarchaeota (Spang et al. 2015; Williams et

al. 2013; Zaremba-Niedzwiedzka et al. 2017). As there is no other evidence to maximally constrain the time of divergence between Eukaryota and Asgardarchaeota, we used the same maximum placed on LUCA; that is, the Moon-forming impact. These key time constraints were combined with nine others (see Results) to calibrate a timescale of life estimated from a dataset of 29 highly conserved, mainly ribosomal, universally distributed proteins (see Methods) using a relaxed molecular clock modelled in a Bayesian framework.

2.2 Methods

2.2.1 Molecular Dataset collation and phylogenetic analysis

The dataset consists of 102 species and 29 universally distributed, protein-coding genes (Table 2.1.). All our data and scripts are available at https://bitbucket.org/bzxdp/ betts_et_al_2017. Proteomes were downloaded from GenBank (Benson et al. 2013) and putative orthologues were identified using BLAST (Altschul et al. 1990). The top hits were compiled and aligned into gene-specific files in MUSCLE (Edgar 2004) and trimmed to remove poorly aligned sites using Trimal (Capella-Gutiérrez, Silla-Martínez, and Gabaldón 2009). To minimize the possible inclusion of paralogues and laterally transferred genes, we generated gene trees (under CAT-GTR + G) in PhyloBayes (Lartillot, Lepage, and Blanquart 2009) and excluded sequences when the tree topology suggested that they might have been paralogues. The sequences were then concatenated into a supermatrix using FASconCAT (Kück and Meusemann 2010), and phylogenetic analyses were performed using PhyloBayes (Lartillot, Lepage, and Blanquart 2009). The superalignment was initially analysed under both GTR + G and CAT-GTR + G (Lartillot and Philippe 2004). RogueNaRok (Aberer, Krompass, and Stamatakis 2012) was used to identify rogue taxa, and analyses were repeated (under both GTR + G and CAT-GTR + G) after unstable taxa were excluded. One final analysis was performed that included only the eukaryotic sequences in our dataset (under CAT-GTR + G). For all PhyloBayes analyses, convergence was tested in PhyloBayes using BPCOMP and TRACECOMP.

Table 2.1. Gene families used in this study by *S. cerevisiae* identification code.

S. cerevisiae gene IDs	Gene family number (arbitrary, corresponds to dm_XX.fa naming scheme in this study)
Rps14bp	(1)
Rps23bp	(6)
Fun12p	(14)
Rpl11ap	(15)
Rsp3p	(20)
Rps16ap	(22)
Rpl1ap	(24)
Rpl2bp	(29)
Rpl23bp	(30)
Rpl12ap	(31)
Eft1p	(33)
Kae1p	(34)
Rps0bp	(35)
Rps2p	(36)
Rps5p	(37)
Srp54p	(40)
Tef1p	(4)
Rli1p	(5)
Dps1p	(10)
Rpa190p	(11)
Sec61p	(12)
Cct5p	(16)
Rfc2p	(17)
Vma2p	(23)
Map2p	(25)
Rpl16ap	(28)
Gln4p	(32)
Rpa135p	(39)
Srp101p	(41)

2.2.2 Calibrations

In total, we used 11 calibrations spread throughout the tree but mainly found within the Eukaryotes as this group has the best fossil record. Calibration choice was carried out conservatively using coherent criteria (Parham et al. 2011). This means that for each calibration the best record will have the following list of things; an up to date phylogenetic analysis, the locality and stratigraphic level from which the fossil / record originates, and, correlation to a published radioisotopic age and/or numeric timescale for example the Geological Time Scale (Gradstein et al. 2012). Therefore, when laying out the calibration a comprehensive assessment of the phylogenetic placement of the fossil based upon morphological characteristics in the first thing needed. Followed by a rationale explaining the minimum date for the

record and, similarly, a maximum date for each calibration. The latter most often formed by information about when that group is no longer in evidence in the rock record. Full details of each calibration are discussed in section 2.3.1.

2.2.3 Divergence time analyses

For our clock analyses, we used a constraint tree based on our CAT-GTR + G and GTR + G trees (Figures 2.1-2.5). The complete phylogeny was rooted to separate Eubacteria from the other lineages (that is, Archaebacteria and Eukaryota) following the topologies generated by phylogenetic analyses detailed in section 2.2.1. To select the amino acid model to be used in our molecular clock analyses, we used PartitionFinder version 1.1.1 (Lanfear et al. 2012). Divergence time estimation was carried out using the approximate likelihood calculation in MCMCTree version 4.9 (Yang 2007). We set four different calibration density distributions: uniform, skewed towards the minimum, skewed towards the maximum and midway between these two dates. For this, we used the Uniform and Cauchy distribution models within MCMCTree, which can be set to place the maximum probability of the node falling in a certain space between the calibrations. The values for these were first produced using MCMCTreeR (https://github.com/ PuttickMacroevolution/MCMCTreeR) code in R (Team 2013). We investigated two strategies to model amino acid sequence evolution: a single WAG + G model or the optimal partitioned model suggested by PartitionFinder. The optimal partitioned model used 29 gene-specific models (28 LG + G and one WAG + G). This means that in MCMCTree each gene was used as a separate partition and assigned its best fitting model. The AIC was used to test whether using a single model or a partitioned model provided a better fit to the data. Rate variation across lineages was modelled using both an autocorrelated and uncorrelated clock model. Bayesian cross-validation was used to test whether one of the two considered, relaxed molecular clock models best fitted the data (implemented in PhyloBayes).

In all our molecular clock analyses using MCMCTree, we applied a soft tail of 2.5% to the upper calibration bound and a hard minimum, apart from the root node (to which a hard maximum was applied) and the nodes calibrated using *Bangiomorpha* (Butterfield, Knoll, and Swett 1990) (to which

a soft minimum tail of 2.5% was applied). For all molecular clock analyses, convergence was tested in Tracer (Rambaut et al. 2018) by comparing plots of estimates from the two independent chains and evaluating whether—for each model parameter and divergence time estimate—the effective sample size was sufficiently large. All reported molecular clock analyses reached excellent levels of convergence.

An assessment of co-estimating time and topology was carried out using MrBayes 3.2.6 (Ronquist, Teslenko, et al. 2012) under the LG model of substitution with a discrete gamma model of rate variation with four bins. A uniform prior was placed on the topology, except for the 10 internal nodes with set time priors which were constrained to be monophyletic. Prior time constraints on these nodes and the root were set as uniform distributions with the bounds taken from the fossil ages – as in all our other analyses. Branch rates were sampled assuming an uncorrelated Independent Gamma Rates (IGR) model (Lepage et al. 2007) with variance sampled from an exponential distribution (mean = 10). The MCMC model sampled every 1000 generations with four independent runs. The tree was summarised as a 50% majority-rule consensus, and model convergence was assessed by analysing Potential Scale Reduction Factor (PSRF, target < 1.05), Effective Sample Size (ESS, target > 200), and visual inspection using TRACER (Rambaut et al. 2018).

2.3 Results

2.3.1 Fossil Calibrations

Node: Last universal common ancestor (LUCA)

Locality and Stratigraphy level: Strelley Pool Formation, Western Australia

Minimum age: 3347 Ma (3350 Ma ± 3 Myr (Nelson 2005))

Maximum age: 4520 Ma (4510 Ma ± 10 Myr (Barboni et al. 2017; Hanan and Tilton 1987))

Phylogenetic justification:

There are numerous reports of fossils from early Archaean sediments, however, determining a biotic origin for these records is difficult. Generally, there is a dearth of strata representative of early Earth history; those strata that are representative and are available for sampling have often been heavily altered by metamorphic processes. The oldest rocks available include, the Itsaq Gneiss, Isua, Greenland; the Barberton Greenstone Belt, South Africa; and the Pilbara Craton, Australia. These contain the oldest possible remains of life. At >3.7 Ga the Itsaq Gneiss contains putative fossils (Dodd et al. 2017; Pflug and Jaeschke-Boyer 1979), stromatolites (Nutman et al. 2016), carbon isotopes (Rosing 1999) and graphite inclusions (Mojzsis et al. 1996; Schidlowski 1988). However, each of these records has been disputed, either considered unlikely to be fossils, or that the record could be produced by geological rather than biological means (Lepland, Arrhenius, and Cornell 2002; Van Zuilen, Lepland, and Arrhenius 2002; van Zuilen et al. 2003) i.e. isotope ratios and graphite inclusions, synthesized by Fisher-Tropsch type (FTT) reactions (Horita and Berndt 1999; Lollar et al. 2002).

At Pilbara, there are claims of isotopic evidence for sulphur bacteria (Shen, Buick, and Canfield 2001), putative stromatolites and the infamous microfossils from the Apex Chert (Schopf 1993), as well as other microfossil reports (Buick 1990; Ueno 2001). None of these records is conclusive, when re-examined the Apex Chert microfossils (Schopf 1993) proved more likely to be an artefact of the reorganization of carbonaceous matter (Brasier et al. 2002; Brasier et al. 2005). Likewise, the other microfossils have not been rigorously examined and so do not provide conclusive evidence of life. The sulphur isotope data (Shen, Buick, and Canfield 2001) is also uncertain as it is possible to produce the

same signals by non-biological means (Runnegar et al. 2001). Microfossils have also been reported from Barberton (Engel et al. 1968; Walsh and Lowe 1985; Westall et al. 2001; Westall et al. 2006) but their biogenesis has been disputed.

Putative stromatolites are widespread in ancient sediments in both Barberton and Pilbara (Allwood et al. 2006; Allwood et al. 2007; Byerly, Lower, and Walsh 1986; Hofmann et al. 1999; Walter, Buick, and Dunlop 1980; Van Kranendonk 2006) but their formation is not exclusively tied to the presence of biological processes and the oldest stromatolites are most often found without any accompanying microbial fossils. Their abiogenic synthesis has been replicated laboratory conditions (McLoughlin, Wilson, and Brasier 2008; Lowe 1994) and so they provide an uncertain record. Therefore, we must look for more conclusive evidence of life, that which has been examined from several angles. More rigorous analysis has been undertaken of fossils from slightly younger sites. For example, a sample of fossils from the ~3.2 Ga Moodies Group, Barberton, were described using criteria which looked at a rigorous range of criteria: fossil placement within the rock; their ultrastructure; their composition; and their size (Javaux, Marshall, and Bekker 2010). Some of these small organic walled fossils are actually very large (up to 300 microns diameter) (Javaux, Marshall, and Bekker 2010); sizes which are unknown amongst archaea (Dworkin 2006). Older remains from the Strelley Pool Formation, Pilbara, Western Australia (Sugitani et al. 2010; Sugitani, Mimura, Takeuchi, Lepot, et al. 2015) have also been examined based on a set of criteria similar to those used by Javaux and colleagues. These fossils have a complex ultrastructure and acid resistant walls that survive being digested out of the rock. Additionally, it should be noted that the organic carbon signature shows that the fossils were not emplaced into the rock at a later stage, a problem with many early records. Some of these fossils are also present in multi-cell chains. These are not known to form in abiotic ways and, hence, it can be concluded that these structures are biological in origin. The Strelley Pool Formation also contains a host of other evidence for life. These include other microfossils both alone (Sugitani et al. 2013) and in association with pyrite crystals (Wacey et al. 2011), possibly indicating some kind of sulphur metabolism backed up a previous study showing sulphur metabolism (Wacey et al. 2010), as well as microbial mats (Duda et al. 2016), and stromatolites, which have been more intensely studied to give credence to their biological affinity (Wacey 2010). What is more the microfossils have been shown to

possess specific $\delta^{13}C_{org}$ signatures that are correlated specifically to the microfossils (Lepot et al. 2013). Overall these show a diverse community (Sugitani, Mimura, Takeuchi, Yamaguchi, et al. 2015). Although alone these would not provide a suitable record, in accordance with the well-studied fossils (Sugitani, Mimura, Takeuchi, Lepot, et al. 2015) they provide a robust calibration with which to constrain LUCA.

Age justification:

Hard minimum: The Strelley Pool Formation is located in North Eastern Australia and is part of the larger Pilbara Craton. The stratigraphic position of this formation (also known as the Strelley Pool Chert) has been contentious but it is now argued to form a layer between the Warawoona and Kelly groups (Hickman 2008). The formation is dated to 3426-3350 Ma (Hickman 2008), with the minimum age (3350 Ma ± 3 Myr) based on a volcaniclastic tuff, at the base of the overlying Euro Basalt (Nelson 2005) in the Kelly Group. Hence our minimum age constraint is 3347 Ma.

Maximum: We can use the Moon-forming impact as a maximum constraint; there is no other event or date of significance which can be used in its place. This devastating event would have sterilised the Earth, hence any life now present on the planet must have evolved post-impact. It has been proposed that life would not have been able to survive the late heavy bombardment, which post-dated the Moon-forming impact, but this view has been contested as ideas of a cool early earth and an early ocean have been proposed (Ryder 2002; Valley et al. 2002), as well as models which show that life would have been able to survive during this intense bombardment (Abramov and Mojzsis 2009). It is also possible that there was no late heavy bombardment because evidence of its occurrence has been found on the Moon but not on Earth (Koeberl 2006). There is some debate over the exact timing of the impact with proposed dates ranging from 4540 Ma ± 10 Myr (Kleine et al. 2005) to ~4440 Ma (Carlson and Lugmair 1988). Some of the most recent simulations and models place the Moon-forming impact at ~4470 Ma based on asteroidal meteorites and siderophile elements (Bottke et al. 2015; Jacobson et al. 2014). This concurs with estimates based on U-Pb isotopes (Tera, Papanastassiou, and Wasserburg 1974), Hf/W isotopes (Halliday et al. 1996) and Rb/Sr isotopes (Halliday 2008). We use the oldest credible date to encompass reasonable uncertainty. The oldest date of 5400 Ma is based on the Hf-W system (Kleine et al. 2005; Kleine et al. 2002), around which there is some debate as to the amount of signal caused by

cosmogenic production of ^{182}W from ^{181}Ta (Touboul et al. 2007). Hence, the most credible date comes from the U-Pb system. We follow other critical reviewers (Halliday 2014) in accepting Pb-Pb dating carried out on Moon rocks, yielding a date of 4510 Ma ± 10 Myr (Hanan and Tilton 1987): a date which has also recently been confirmed by reanalysis of the Apollo zircons (Barboni et al. 2017). Thus, our maximum constraint is 4520 Ma.

Node: Total group Cyanobacteria

Locality and Stratigraphy level: Manzimnyama Banded Ironstone Formation, Fig Tree Group, Barberton, South Africa

Minimum age: 3225 Ma (3226 Ma ± 1 Myr (Kamo and Davis 1994))

Maximum age: 4520 Ma (4510 Ma ± 10 Myr (Hanan and Tilton 1987))

Phylogenetic justification: Cyanobacteria are the only living group of organisms that have evolved oxygenic photosynthesis. Proposed records of Cyanobacteria from ancient rocks include Banded Ironstone Formations (BIFs), stromatolites, biomarkers, and a number of isotope systems. BIFs, which are found among the oldest sedimentary rocks, including protoliths of the 3.8 Ga Itsaq Gneiss, show a reduction of ferrous iron which has been claimed to occur due to cyanobacterial effects. However, arguments have been presented for the production of BIFs via abiogenic ultra-violet induced photolysis (Cairns-Smith 1978) and anoxygenic bacterial photosynthesis (Crowe et al. 2008; Konhauser et al. 2002). Early stromatolites are not sufficient evidence as they are not all biogenic and they don't necessarily require Cyanobacteria for formation (Grotzinger and Rothman 1996; McLoughlin, Wilson, and Brasier 2008). The best indicator of free oxygen at levels incompatible with photolysis, is from isotopes. These are a good proxy for oxygen because many elements are very sensitive to oxidative weathering. Prior to the Great Oxygenation Event, oxygen records in the form of isotopes extend back to 3.25 Ga (Satkoski et al. 2015). The authors report stable Fe and U-Th-Pb isotopes from the Manzimnyama BIF in the Fig Tree Group, Barberton, South Africa, which indicate a level of free oxygen indicative of cyanobacterial activity. They also find that there is a stratification in oxygen levels at the site, showing an oxygenated shallow water layer and an anoxic deeper water. They argue that this is what we would expect to see in areas where there is some cyanobacterial activity. It is possible that

oxygen was being produced in smaller quantities prior to the GOE and that these pockets of oxygen could be concentrated in an otherwise anoxic water column (Olson, Kump, and Kasting 2013). Other evidence for oxygenation from within this sequence comes from the Moodies group which lies immediately above the Fig Tree Group at Barberton. This has macroscopic tufted microbial mats (Homann et al. 2015), that are thought to grow upwards towards a source of light, and in modern examples are made mostly of cyanobacteria. Additionally, this evidence for oxygenation is not isolated as numerous other lines of evidence, based mainly upon redox sensitive elements and other isotopes, now support the appearance of pre-GOE oxygen being produced by cyanobacteria (Anbar et al. 2007; Crowe et al. 2013; Kendall et al. 2010; Czaja et al. 2012; Planavsky et al. 2014; Riding, Fralick, and Liang 2014).

Age justification:

Hard minimum: The isotopic evidence from the Manzimnyama BIF in the Fig Tree Group, Barberton, South Africa (Satkoski et al. 2015). The age of the Fig Tree Group is well constrained with a spherule layer at its base dated at 3258 Ma ± 3 Myr (Byerly et al. 1996), and an overlying volcanic unit at its top dated at 3226 Ma ± 1 Myr (Kamo and Davis 1994). Hence, the minimum date we would assign is 3225 Ma.

Maximum: See Maximum for LUCA node.

Node: Total group Eukarya

Locality and Stratigraphy level: Changcheng Group, Hebei Province, North China

Minimum age: 1619.1 Ma (1625.3 ± 6.2 Myr (Li et al. 2013))

Maximum age: 4520 Ma (4510 Ma ± 10 Myr (Hanan and Tilton 1987))

Phylogenetic justification:

The record of eukaryotes covers a large timespan, during much of which the fossils attributed to eukaryotes are relatively simple and do not exhibit much morphological variation. The earliest of these that have been rigorously examined are those from the Changcheng Group in North China. These fossils come from two levels within this group, the Changzhougou Fm. and the Chuanlinggou Fm (Lamb et al. 2009; Peng, Bao, and Yuan 2009). The units are made up of sandstone and shale, within which the

fossils are found. The fossils are small and lenticular in shape with a carbonaceous outer sheath and what are interpreted to be excystment structures. The complexity exhibited by these sheaths and the inferred function, along with the size, places them into the eukaryote domain. The forms preserved at Changcheng are large enough, on average >125μm that they unlikely to be any kind of Eubacteria or Archaebacteria. Some bacterial cells can reach large sizes and size is not the best criteria to use but can be informative when used in conjunction with other characteristics. The authors demonstrate that the cells have a double sheath. The possibility that cyanobacteria have these structures is discussed but refuted on the basis of size. They are even proposed to be part of the green-algae plant lineage (Moczydłowska et al. 2011). However, it is due to a lack of definitive features this claim cannot be substantiated. The age of these fossils encompasses reports of other fossils that are also Eukaryotic in nature, but those which also have uncertain affinities, such as the probable 1.56 Ga multicellular fossils (Zhu et al. 2016), the string of beads *Horodyskia* (Horodyski 1982), and *Shuiyousphaeridium* (Yin 1997) and other acritarch and leiosphaerid forms (Knoll et al. 2006; Cohen and Macdonald 2015). Unfortunately, these fossils are not diagnostic of any crown group eukaryotes and so we can only use them to calibrate the total group of eukaryotes, helping us to provide a robust minimum for their appearance. Putative rhodophytes from the Chitrakoot Formation are slightly younger (see total-group Rhodophyta, below). Although some are sceptical of the eukaryotic nature of these fossils (Knoll 2014), the combination of their morphology and size seems sufficient to assign them to a stem group eukaryote affinity.

Age justification:

Hard minimum:

As the oldest of these fossils are found in the Changzhougou Formation it is this that we can date. To acquire a minimum date for the whole formation, we use ash layers in the overlying formation, yielding an age of 1625.3 ± 6.2 Myr (Li et al. 2013). The microfossils are present in both these layers but have been described separately (Lamb et al. 2009; Peng, Bao, and Yuan 2009). Hence, we can use the date of the oldest Chuanlinggou, 1619.1 Ma, to date the underlying Changzhougou.

Maximum: See Maximum for LUCA node.

Node: Total group Rhodophyta

Specimen and fossil taxon: *Bangiomorpha pubescens.* (Holotype) HUPC 62912, Slide HUST-1A, England Finder coordinates: O-35.

Locality and Stratigraphy level: Lower Hunting Formation, Somerset Island, arctic Canada.

Soft Minimum age: 1030 Ma (1092 Ma ± 59 Myr (Gibson et al. 2017))

Soft Maximum age: 1891 Ma (1823 Ma ± 68 Myr (Lu, Yang, and Zhu 1996))

Phylogenetic justification: There are several reports of red algae within the fossil record, stretching back into the Ediacaran, Neo- and Meso-proterozoic. The oldest of which are 1600 million year old fossils, *Rafatazmia chitrakootia* and *Ramathallus lobatus*, from the Chitrakoot Formation (Bengtson, Sallstedt, et al. 2017). However, though both are suggested to be red algae and, while the remains are compatible with this interpretation, they no not preclude alternative assignments within total group Archaeplastida. *Bangiomorpha pubescens* is younger fossil, originally described as a Bangiale red algae in comparison to the extant *Bangia* due to the distinctive, radially orientated, intercalary division of its cells and its putative development (Butterfield, Knoll, and Swett 1990; Butterfield 2000). It has therefore been used as a calibration for the red algae or sometimes more specifically for the bangiophyte red algae (Eme et al. 2014; Parfrey et al. 2011). Red algae are united by general characteristics that are not commonly preserved in the fossil record, even in the most exceptional of circumstances, e.g. the red coloured pigments, and unstacked thylakoids within the chloroplasts (Hoek et al. 1995; Lee 2008). Hence, *Bangiomorpha* was identified using potential developmental characters and the distinct shape of its cell arrangements. However, although these characters are distinctive (Hoek et al. 1995), they are also characteristic of several other red algae (Yang et al. 2016). *Bangiomorpha* has been described as having a multicellular holdfast, a feature found in some Compsopogonophyceae, another group of basal red algae. Modern *Bangia* has an attachment rhizoid, not a multicellular holdfast indicating that the features of *Bangiomorpha* are not specifically Bangiale. These observations make it inappropriate to assign *Bangiomorpha* specifically to Bangiales. However, the distinct developmental, reproductive and morphological characteristics appear sufficient to assign *Bangiomorpha* to Rhodophyta as a whole.

Hence, we can use this fossil to calibrate the node subtending Rhodophyta which link them to their nearest common ancestor.

Age justification:

Soft minimum constraint: The oldest records of *Bangiomorpha pubescens* occur in the Lower Hunting Formation, of Somerset Island, Arctic Canada. A minimum age for the formation is based on the age of the Franklin igneous events, which have been dated to 723 Ma ± 3 Myr (Heaman, LeCheminant, and Rainbird 1992), with a maximum age of 1267 Ma ± 2 Myr based on the McKenzie igneous events (LeCheminant and Heaman 1989). The original description (Butterfield, Knoll, and Swett 1990) cites an unpublished Pb-Pb date 1198 Ma ± 24 Myr as a best date for *B. pubescens*, however, this date remains unsubstantiated and so it must be discounted. The formation from which *Bangiomorpha* was recovered can be correlated lithostratigraphically to the Society Cliffs Formation (Kah et al. 1999) and the Uluskan Group (Mayr 2004), which are closer to the base of the sequence, and dated at ~1267 Ma (Mesoproterozoic). This is substantially older than the ~723 Ma minimum constraint on the age of the Lower Hunting Formation. The other option is a date of 1092 ± 59 Myr (Turner and Kamber 2012) established from a shale layer present in the Arctic Bay formation, which is comparable (Long and Turner 2012) to the sequences below the *Bangiomorpha* fossiliferous layer i.e. the Lower Hunting formation. However, this date is older than the layer in which *Bangiomorpha* resides. The age of the fossil *Bangiomorpha* can now be more precisely dated to 1047 +13/–17 Ma (a minimum of 1030 Ma) (Gibson et al. 2017). Hence, we use a hard minimum of 1030 Ma.

Soft Maximum Constraint: The soft maximum constraint is based on the earliest record of eukaryotes (Lamb et al. 2009; Peng, Bao, and Yuan 2009; Zhongying 1986) when, despite the presence of simple eukaryotes, there is no evidence of anything as complex as a definitively multicellular alga. Though the fossils present have been suggested by some to represent some kind of green algae (Moczydłowska et al. 2011). The maximum for this formation is based on the igneous and metamorphic rocks that it overlies. These rocks are dated at 1823 Ma ± 68 Myr (Lu, Yang, and Zhu 1996), yielding a soft maximum constraint of 1891 Ma.

NOTE: The Gibson et al., 2017 paper was not available when the analyses for Chapter 2 were carried out. Hence, although Chapters 3 and 4 employ this calibration, Chapter 2 used a soft minimum of 2.5%

and a date of 1033 Ma. This was done using the date of 1092 ± 59 Myr (Turner and Kamber 2012) established from a shale layer present in the Arctic Bay formation, which is comparable (Long and Turner 2012) to the sequences below the *Bangiomorpha* fossiliferous layer i.e. the Lower Hunting formation. Due to this layer being slightly above the one containing *Bangiomorpha* an approach using a soft minimum was employed.

Nodes: Crown Alphaproteobacteria

Specimen and fossil taxon: *Bangiomorpha pubescens.* (Holotype) HUPC 62912, Slide HUST-1A, England Finder coordinates: O-35.

Locality and Stratigraphy level: Lower Hunting Formation, Somerset Island, arctic Canada.

Soft Minimum age: 1030 Ma (1092 Ma ± 59 Myr (Gibson et al. 2017))

Soft Maximum age: 4520 Ma (4510 Ma ± 10 Myr (Hanan and Tilton 1987))

Phylogenetic justification: There are no fossils that can be attributed to Alphaproteobacteria. However, the important eukaryote organelle, the mitochondria has been found by consensus to have belonged within Alphaproteobacteria. This is because mitochondria formed via an endosymbiosis event with the protoeukaryote (Roger, Muñoz-Gómez, and Kamikawa 2017). Within the Alphaproteobacteria group the mitochondrion are most commonly linked to the Rickettsiales (Williams, Sobral, and Dickerman 2007; Wang and Wu 2015) though arguments have also been made for them belonging to other alphaproteobacterial groups (Roger, Muñoz-Gómez, and Kamikawa 2017; Atteia et al. 2009; Esser et al. 2004). Mitochondria contain a mosaic of genes which are not all alphaproteobacterial in origin (Gray 2015, 2012), but nonetheless it is still believed to have originated within this group. *Bangiomorpha pubescens* (Butterfield, Knoll, and Swett 1990) is a total group rhodophyte with features that link it to the basal rhodophyte groups such as its cell arrangement, and others which mean it cannot be placed specifically within any one of them. It is the oldest fossil in the record that can be confidently identified as a crown-eukaryote. There are older fossils that are eukaryotic in nature, but they cannot be placed with certainty into crown-Eukaryota. Hence, we can use *Bangiomorpha* to provide some level of constraint to the Alphaproteobacteria, in a part of the tree of life that is otherwise poorly constrained.

Age justification:

Soft minimum constraint: See Total group Rhodophyta soft minimum constraint.

Maximum: See Maximum for LUCA node.

NOTE: The Gibson et al., 2017 paper was not available when the analyses for Chapter 2 were carried out. Hence, although Chapters 3 and 4 employ this calibration, Chapter 2 used a soft minimum of 2.5% and a date of 1033 Ma. This was done using the date of 1092 ± 59 Myr (Turner and Kamber 2012) established from a shale layer present in the Arctic Bay formation, which is comparable (Long and Turner 2012) to the sequences below the *Bangiomorpha* fossiliferous layer i.e. the Lower Hunting formation. Due to this layer being slightly above the one containing *Bangiomorpha* an approach using a soft minimum was employed.

Nodes: Crown-Cyanobacteria

Specimen and fossil taxon: *Bangiomorpha pubescens.* (Holotype) HUPC 62912, Slide HUST-1A, England Finder coordinates: O-35.

Locality and Stratigraphy level: Lower Hunting Formation, Somerset Island, arctic Canada.

Soft Minimum age: 1030 Ma (1092 Ma ± 59 Myr (Gibson et al. 2017))

Soft Maximum age: 4520 Ma (4510 Ma ± 10 Myr (Hanan and Tilton 1987))

Phylogenetic justification: Cyanobacteria are inferred to have a relatively plentiful fossil record. Often the Great Oxidation Event (GOE) and a number of fossils are used to calibrate the origins of the crown group and various lineages within it. However, the assumption that the GOE was caused by crown Cyanobacteria rests on the assumption that photosynthesis evolved in associated with the crown clade. This has been recently challenged and so we do not use it as a calibration here (Shih and Matzke 2013). Potential records of Cyanobacteria extend into the Archaean, but these are mainly simple cells and filaments (Schopf 2006) whose affinities cannot be substantiated (Brasier et al. 2006). There are fossils described as akinetes, cyanobacterial resting spores, from 2100 Ma and 1600 Ma (Tomitani et al. 2006). However, modern specimens show a range of characters and morphology making it difficult to relate these to any potential ancient counterparts, and other bacterial cells can also show this type of simple morphology (Butterfield 2015b). The most convincing fossil remains are found in the Belcher Formation, Canada (Golubic and Hofmann 1976; Hofmann 1976), from around 1.9 billion years old,

however, even these cannot be discriminated confidently from other bacterial grades (Butterfield 2015b). Instead of using the above-mentioned fossils as calibration points, as in other studies (Sánchez-Baracaldo et al. 2017), we opted for a more conservative approached and used evidence for the oldest archaeplastid; this would have had a chloroplast, known to have originated in an endosymbiotic event with a Cyanobacteria. There is no strict consensus as to which cyanobacterial group plastids evolved from with the main argument being whether they evolved from an early (Ponce-Toledo et al. 2017) or late (Ochoa de Alda et al. 2014; Deusch et al. 2008) branching lineage within Cyanobacteria. *Bangiomorpha pubescens* (Butterfield, Knoll, and Swett 1990) is a total group Rhodophyte (see total-group Rhodophyta, above). It is the oldest fossil in the record that can be confidently identified as a crown group eukaryote; there are older fossils that are eukaryotic in nature, but they cannot be placed with any certainty into one of the extant eukaryotic groupings.

Age justification:

Soft minimum constraint: See Total group Rhodophyta soft minimum constraint.

Maximum: See Maximum for LUCA node.

NOTE: The Gibson et al., 2017 paper was not available when the analyses for Chapter 2 were carried out. Hence, although Chapters 3 and 4 employ this calibration, Chapter 2 used a soft minimum of 2.5% and a date of 1033 Ma. This was done using the date of 1092 ± 59 Myr (Turner and Kamber 2012) established from a shale layer present in the Arctic Bay formation, which is comparable (Long and Turner 2012) to the sequences below the *Bangiomorpha* fossiliferous layer i.e. the Lower Hunting formation. Due to this layer being slightly above the one containing *Bangiomorpha* an approach using a soft minimum was employed.

Node: Dikarya

Locality and stratigraphy level: Rhynie, Aberdeenshire, Scotland. Lower Devonian

Minimum age: 392.1 Ma (393.3 Ma ± 1.2 Myr (Gradstein et al. 2012))

Maximum age: 1891 Ma (1823 Ma ± 68 Myr (Lu, Yang, and Zhu 1996))

Phylogenetic justification: The minimum constraint is based upon fossils from the Rhynie Chert (Taylor, Hass, and Kerp 1999) described as *Paleopyrenomycites devonicus* (Taylor et al. 2005). This

fungal fossil is found in association with the roots of early plants and has key characteristics that relate it to the Ascomycota, including containing the sexual spores (asci) in a sac-like structure, the ascus. Although there are earlier examples of possible fossil fungi much of their interpretation is spurious. This category includes *Tappania*, which was once interpreted as a fungus (Butterfield 2005), but is now considered to be an acritarch (Butterfield 2015a), and the 'lichen-like' fossil from Doushantuo (Yuan, Xiao, and Taylor 2005) is difficult to discriminate from diagenetic artefacts that are characteristic of fossils from the Weng'an Biota (Cunningham et al. 2012). There is a more convincing record of a possible Glomeromycota fungus from the Ordovician (Redecker, Kodner, and Graham 2000). However, this specimen has not been assigned with as much confidence to a distinct fungal lineage as those fossils contained in the younger Devonian Rhynie Chert deposits. The oldest report of a fungi-like fossil is from the Ongeluk Formation, ~2400 Ma (Bengtson, Rasmussen, et al. 2017). The filaments are situated within basaltic lavas, a rock type shown to host putative fungal species in more recent Eocene basalts (Schumann et al. 2004; Ivarsson et al. 2013; Ivarsson et al. 2012). However, although the Ongeluk fossils do show many typical fungal features, these can also be attributed to the actinobacteria, such as the hyphae-like cells and Y-junctions, thus, their affinities are ambiguous. Hence, we use the confidently assigned fungi fossil from the Rhynie chert to constrain the minimum age of the clade comprising Ascomycota and Basidiomycota and sister lineage Glomeromycota.

Age justification:

Hard minimum: Proposed dates for the Rhynie Chert system have been mostly based upon zircons from volcanic deposits in the sequence. Two recent dates proposed are 407.1 Ma ± 2.2 Myr (Mark et al. 2011) and 411.5 Ma ± 1.3 Myr (Parry et al. 2011). The former is from a hydrothermally produced layer within the sequence and with which there is high oxygen isotopic homogeneity from the layers with the spore bearing assemblage (Mark et al. 2011). The other date is derived from the Milton of Noth andesite (Parry et al. 2011). Despite being based on zircon evidence, neither of these dates is suitable; the Milton of Noth andesite has uncertain placement within the sequence but is most likely found beneath the Rhynie spore-bearing layer (Rice and Ashcroft 2003) and so cannot be used to provide a minimum date. The later date (Parry et al. 2013) is also unsuitable because the layers which are dated do not come from above the spore assemblage, and the method of dating has some problems. Therefore,

we base our minimum clade age constraint on the spore assemblage characterizing the Rhynie Chert. This places the Rhynie Chert in the early Pragian to early Emsian (Wellman 2006). The age of the top of the Emsian-Eifelian boundary is dated as 393.3 Ma ± 1.2 Myr (Gradstein et al. 2012). Hence our minimum clade age constraint is 392.1 Ma.

Soft maximum: The maximum for this calibration is based on the earliest record of eukaryotes (Lamb et al. 2009; Peng, Bao, and Yuan 2009; Zhongying 1986) when, despite the presence of simple eukaryotes, there is no evidence of anything as complex as a multicellular alga. Though the fossils present have been suggested by some to represent some kind of green algae (Moczydłowska et al. 2011). This date also encompasses the recent discovery of possible multicellular eukaryotes from the 1560 Ma (Zhu et al. 2016; Rambaut et al. 2018). The maximum for this formation is based on the igneous and metamorphic rocks that lie beneath it. These rocks are dated at 1823 Ma ± 68 Myr (Lu, Yang, and Zhu 1996), thus, our maximum is 1891 Ma.

Node: Crown group Foraminifera

Locality and Stratigraphy level: The Chapel Island Formation, Newfoundland, Canada. Lower Cambrian.

Minimum age: 525.5 Ma (525.5 Myr (Gradstein et al. 2012))

Maximum age: 1891 Ma (1823 Ma ± 68 Myr (Lu, Yang, and Zhu 1996))

Phylogenetic justification:

The foraminifera are a group of testate eukaryotes that are part of Rhizaria, a group that also includes Cercozoa and Radiolaria. Foraminifera are well known from most of the Proterozoic before which there are scattered reports with varying degrees of validity. The very oldest possible reports come from Post-Sturtian deposits located in Namibia and Mongolia (Bosak et al. 2012; Bosak, Lahr, et al. 2011). These are interpreted as foraminifera based on the composition of the tests found. However, the authors cautiously interpret them as foraminifera, partly due to the shape that is not seen in modern forms, so there is still a level of uncertainty in their affinity. Other Ediacaran fossils have been described as foraminifera, such as the enigmatic *Palaeopascichnus*. However, these fossils lack a number of key diagnostic features of foraminifera (Antcliffe, Gooday, and Brasier 2011). Generally, the oldest forms

are regarded to be those from Western African (Culver 1991) and from the Lower Cambrian of Canada (McIlroy, Green, and Brasier 2001). Though Culver described the Western African forms as Cambrian in nature, due to their position and the appearance of a Cambrian snail in the same deposits, new dating suggests that the formation might actually be closer to the Ordovician in age (Villeneuve et al. 2014). The fossil described as *Platysolenites cooperi* (McIlroy, Green, and Brasier 2001) has had its foraminiferal affinity questioned based on the possible composition of their tests (Rozanov 1983; Rozanov et al. 1992). However, in their paper McIlroy and colleagues dispel this doubt by looking in detail at the wall composition. They find that it is composed of agglutinated grains, was organically bound and probably flexible in life. They also find that it shows evidence of fracturing that was repaired during the organism's lifetime, on the outside of the wall, a character not seen in metazoans. This and other support from previous reviews (Lipps 1992; Lipps and Rozanov 1996) provides strong evidence for *P. cooperi* being an early agglutinating foraminifera species.

Age justification:

Minimum: The oldest fossils of *P. cooperi* come from the latest Ediacaran to Lower Cambrian in Newfoundland, the Chapel Island formation (McIlroy, Green, and Brasier 2001). This formation sits just above the Cambrian boundary and is correlated to the Nemakit-Daldyian which has a minimum date of 525.5 Ma according to the latest version of the geological timescale (Gradstein et al. 2012).

Maximum: The maximum for this calibration is based on the earliest record of eukaryotes (Lamb et al. 2009; Peng, Bao, and Yuan 2009; Zhongying 1986) when, despite the presence of simple eukaryotes, there is no evidence of anything as complex as a multicellular alga. Though the fossils present have been suggested by some to represent some kind of green algae (Moczydłowska et al. 2011). This date also encompasses the recent discovery of possible multicellular eukaryotes from the 1.56 Ga (Zhu et al. 2016; Rambaut et al. 2018). The maximum for this formation is based on the igneous and metamorphic rocks that lie beneath it. These rocks are dated at 1823 Ma ± 68 Myr (Lu, Yang, and Zhu 1996), thus, our maximum is 1891 Ma.

Node: Embryophytes

Locality and Stratigraphy level: Qusaiba-1 core from the Quasim formation of northern Saudi Arabia

Minimum age: 448.5 Ma (Clark and Donoghue 2017)

Maximum age: 509 Ma (Clark and Donoghue 2017)

Age justification:

The oldest evidence of embryophytes are trilete spores. We follow Clark and Donoghue (Clark and Donoghue 2017) in dating these to a minimum date of 448.5 Ma. The maximum is placed at the Bright Angel Shale which has a date of 507.2-509 Ma, hence, the maximum that we use to 509 Ma.

Node: Angiospermae

Locality and Stratigraphy level: Cowleaze Chine Member of the Vectis Formation of the Isle of Wight

Minimum age: 125.9 Ma (126.3 Ma ± 0.4 Myr (Clark and Donoghue 2017))

Maximum age: 247.3 Ma (247.1 Ma ± 0.2 Myr (Clark and Donoghue 2017))

Age justification:

The oldest evidence of angiosperms is tricopolate pollen. We follow Clark and Donoghue (Clark and Donoghue 2017) and date the pollen to the Cowleaze Chine Member, Isle of White. This yields a minimum date of 126.3 ± 0.4 Myr and a maximum date of 247.1 Ma ± 0.2 Myr from a rock layer free of angiosperm pollen.

Node: Metazoa

Locality and Stratigraphy level: White Sea Formation, Russia

Minimum age: 550.25 Ma (552.85 Ma ± 2.6 Myr (Benton et al. 2015))

Maximum age: 833 Ma (827 Ma ± 6 Myr (Benton et al. 2015))

Age justification: The oldest uncontroversial evidence for Metazoa is the fossil *Kimberella quadrata*. The oldest specimen of this is found in the White Sea, Russia, for which a minimum date of 552.85 Ma ± 2.6 Myr has been established. The maximum is set as 827 Ma ± 6 Myr from the age of the Bitter

Springs formation that Benton et al., 2015 summarise as showing no evidence of any total group metazoans.

2.3.2 Topological results

We performed phylogenetic analyses of our complete dataset to evaluate whether it supported generally agreed relationships. While the scope of this study is not that of resolving relationships at the root of the tree of life, this is important to make sure that the genes we selected are informative and do not display obvious paralogy or xenology problems. Analyses of the complete dataset failed to converge under both GTR+G and CAT-GTR+G. Irrespective of that the trees inferred under both models reflect current consensus relatively well. CAT-GTR+G analyses in particular invariably found support for the Eocyte tree, even if with *Korarchaeum cryptofilum* as the sister of Eukaryota rather than the Lokiarchaeota (Fig. 2.1). Differently, GTR+G analyses found support for either the Eocyte tree (still with *Korarchaeum cryptofilum* as sister of the Eukaryota) or for Woese's Three Domains Tree (Fig. 2.2a and 2.2b). RogueNaRok (Aberer, Krompass, and Stamatakis 2012) identified five rogue taxa in the dataset (*Korarchaeum cryptofilum, Treponema pallidum, Fibrobacter succinogenes, Cyanophora paradoxa* and *Actinomadura madurae*). CAT-GTR+G analyses performed after excluding these taxa still failed to converge (Fig. 2.3). However, with the exclusion of the relationships among the eukaryotic supergroups, all key relationships in the CAT-GTR+G tree of Figure 2.3 are resolved according to common knowledge. The GTR+G analysis of the RogueNaRok reduced dataset (Fig. 2.4), converged well and resolved the tree in essential agreement with the CAT-GTR+G analysis, supporting in particular the Lokiarchaeota as the sister of the Eukaryota. Overall, these results indicate that instability is limited to the tip-ward part of the tree and this is not unsurprising given that we specifically targeted highly conserved genes to better date the history of life closer to the root rather than the tips. The only area in which our converged GTR+G tree, and our unconverged CAT-GTR+G, tree disagreed with the current consensus were the relationships of the eukaryotic supergroups. This might indicate Long Branch Attraction Artifacts. To test this hypothesis, we performed a CAT-GTR+G analysis including only the eukaryotic taxa and found relationships that are fully compatible with the current consensus (Fig. 2.5). These data were aligned with MUSCLE and trimmed using TrimAl. This indicates

that the eukaryotic relationships in Figure 2.3 and 2.4 probably represent tree reconstruction artefacts caused by the attraction between eukaryotes lineages (like the secondarily amitocondriate *Giardia lamblia*) and the prokaryotes. Accordingly, for our clock analyses we used a fixed tree topology compatible with the trees in Figure 2.3 and 2.4, but where the eukaryotes were resolved as in Figure 2.5 and unstable taxa identified by RogueNaRock (Aberer, Krompass, and Stamatakis 2012) were reintroduced and resolved according current consensus. This tree forms the topology seen in Fig. 2.14.

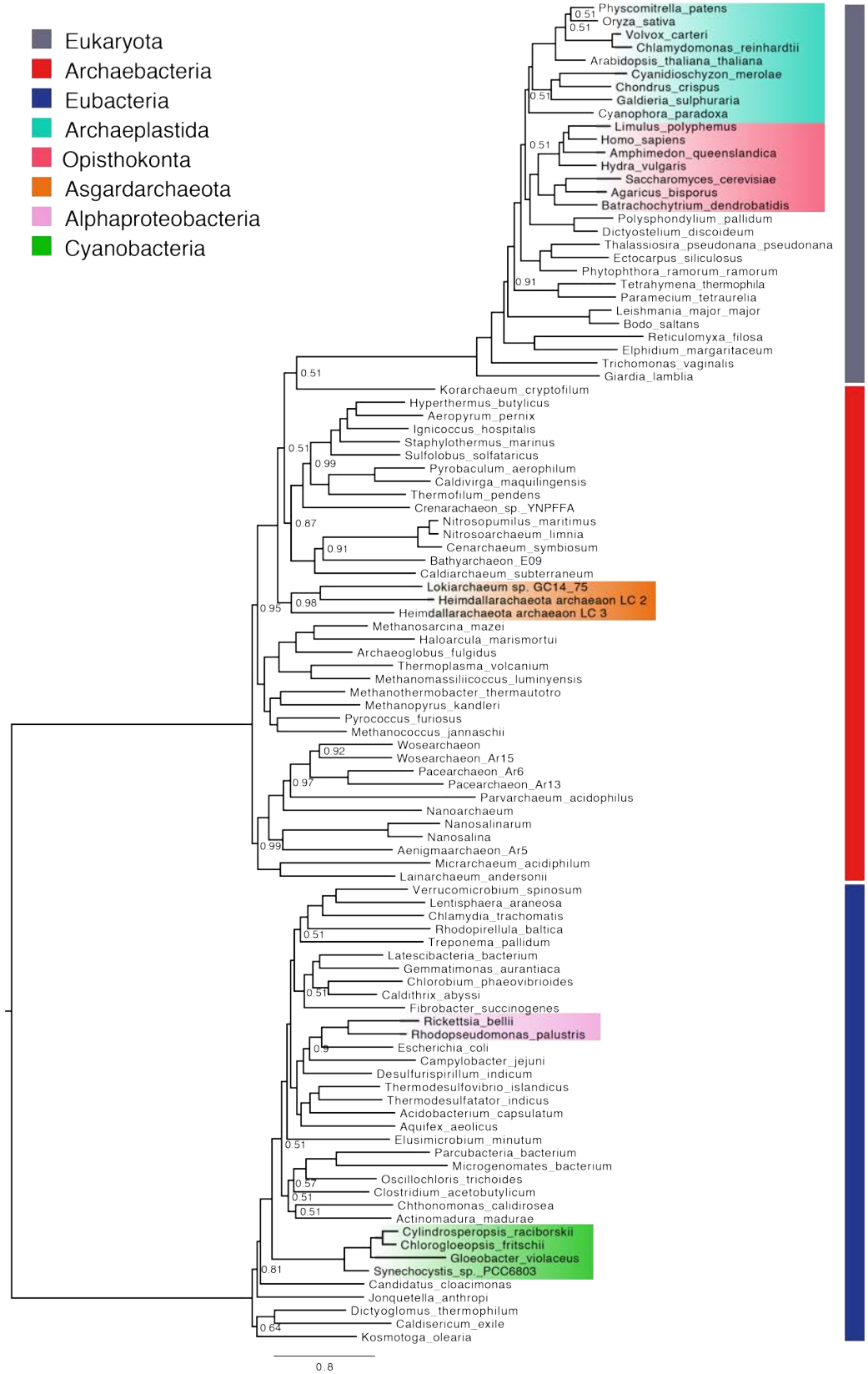


Figure 2.1. Phylogeny produced using PhyloBayes with a CAT-GTR+G model (not converged and including rogue taxa). The numbers within the phylogeny indicate the posterior probability of the node if it is less than 1. The scale at the bottom represents the number of substitutions per site, in this case amino acids. Some lineages have been highlighted.

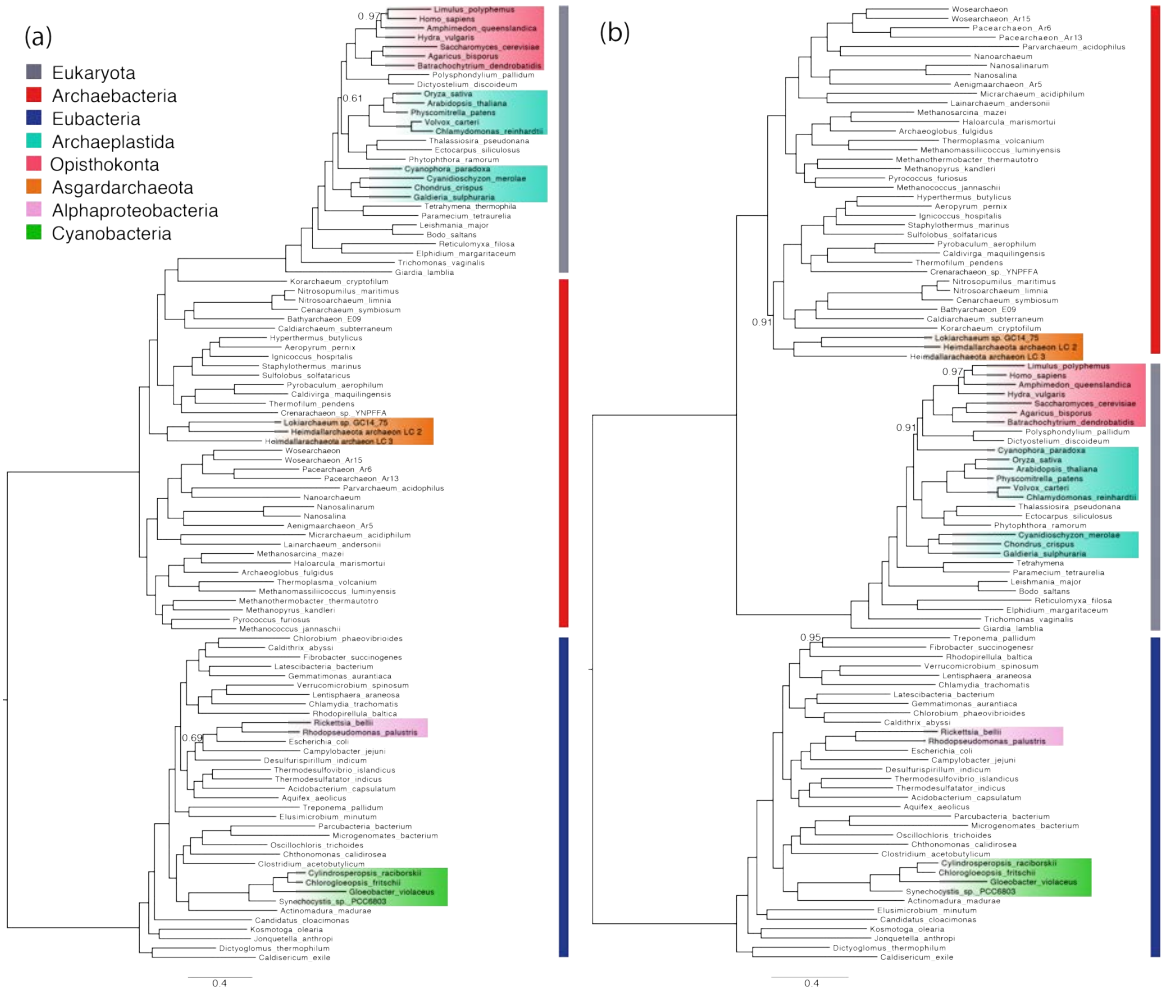


Figure 2.2. Phylogenies produced by two independent runs using PhyloBayes with a GTR+G model (not converged and including rogue taxa) (a) Showing support for the Eocyte tree and (b) for Woese's Three Domains Tree. The numbers within the phylogeny indicate the posterior probability of the node if it is less than 1. The scale at the bottom represents the number of substitutions per site, in this case amino acids. Some nodes of interest have been highlighted.

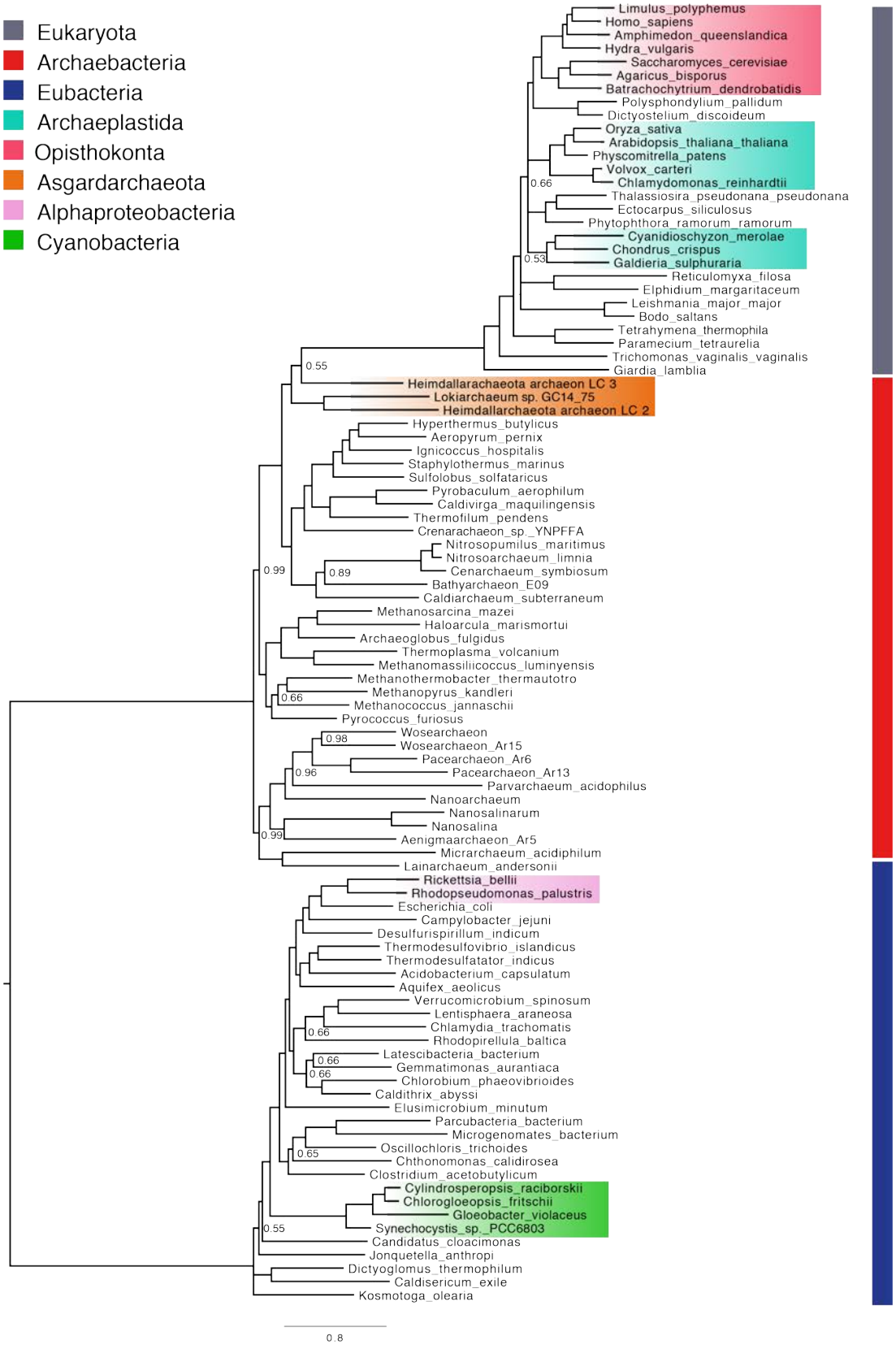


Figure 2.3. Phylogeny produced using PhyloBayes with a CAT-GTR+G model (not converged and excluding rogue taxa). The numbers within the phylogeny indicate the posterior probability of the node if it is less than 1. The scale at the bottom represents the number of substitutions per site, in this case amino acids. Some nodes of interest have been highlighted.

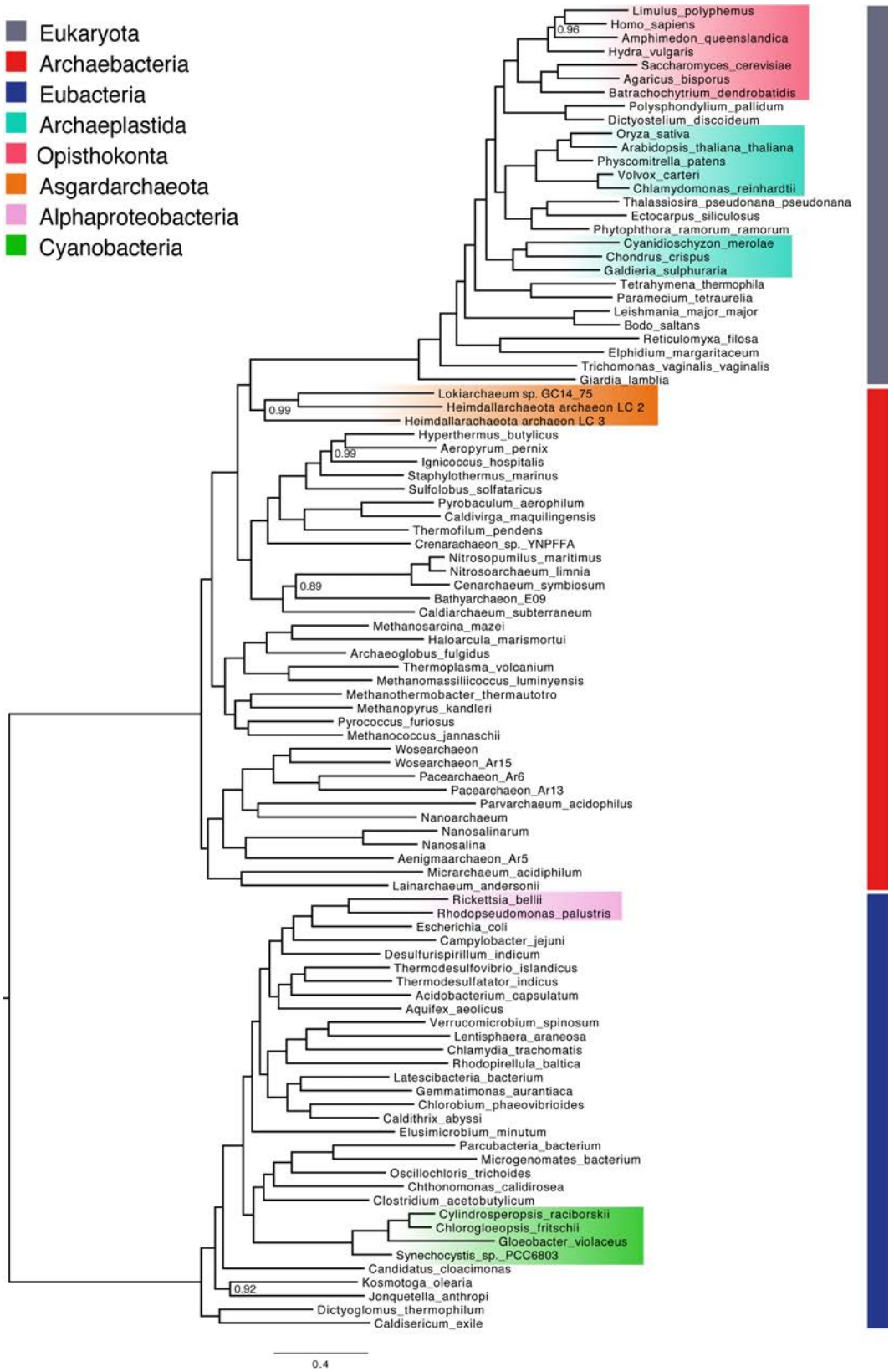


Figure 2.4. Phylogeny produced using PhyloBayes with a GTR+G model. This analysis converged well (number of cycles = 3872; Burnin = 1000; BPcomp Maxdiff = 0.18; Tracecomp Minimum Effective Size = 244; Tracecomp maximum relative difference = 0.15). The numbers within the phylogeny indicate the posterior probability of the node if it is less than 1. The scale at the bottom represents the number of substitutions per site, in this case amino acids. Some nodes of interest have been highlighted.

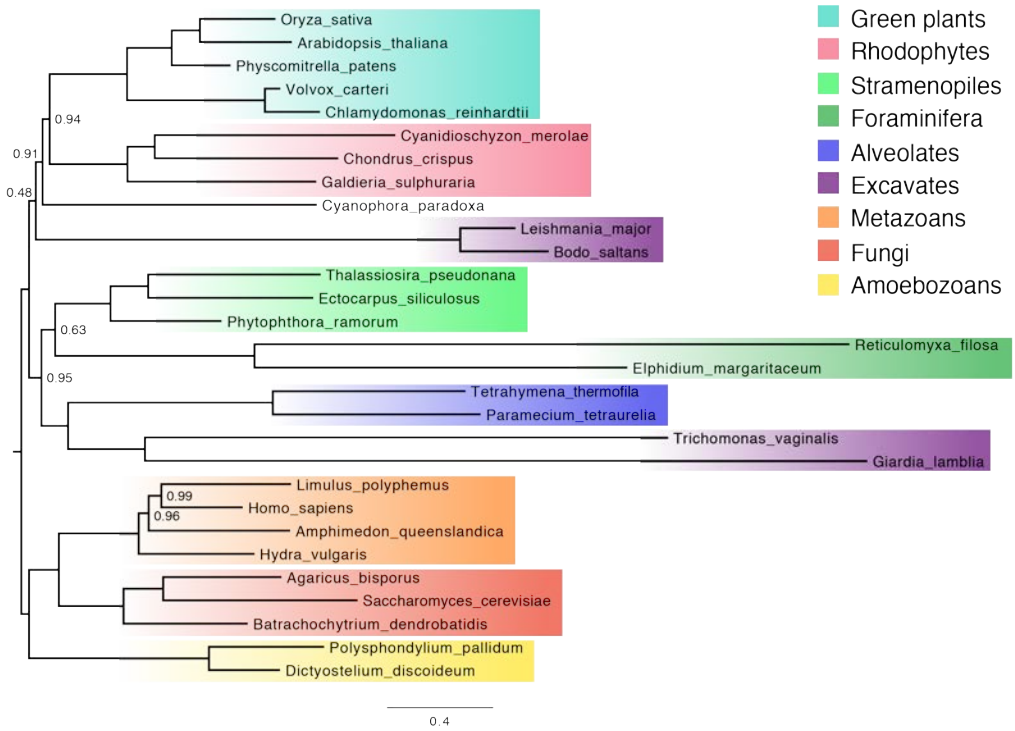


Figure 2.5. Phylogeny showing the Eukaryote only relationships. Produced using PhyloBayes with a CAT-GTR+G model. This analysis reached an acceptable level of convergence (number of cycles = 34660; Burnin = 15000; BPcomp Maxdiff = 0.05; Tracecomp Minimum Effective Size = 170; Tracecomp maximum relative difference = 2.2). The numbers within the phylogeny indicate the posterior probability of the node if it is less than 1. The scale at the bottom represents the number of substitutions per site, in this case amino acids. Some nodes of interest have been highlighted.

2.3.3 Divergence time results

Analytical choices can deeply affect molecular clock posterior age estimates (Warnock, Yang, and Donoghue 2012) and we explored a range of prior probability distributions to model our fossil calibrations and estimate conservative credibility intervals for our divergence times. Initially, we applied a hard maximum of 4.52 Ga (the age of the Moon-forming impact) to the root of our tree and used uniform age priors (reflecting agnosticism about divergence timing relative to constraints) to the other fossil calibrations (Fig. 2.6a). These analyses assumed an uncorrelated molecular clock model and produced the amino acid substitution processes using optimal gene-specific substitution models. Subsequently, we explored the impact of using calibration protocols based on non-uniform age priors. First, we implemented a truncated Cauchy distribution with the mode located halfway between the minimum and maximum bounds, reflecting a prior view that true divergence times should fall between

the minimum and maxi- mum calibration points (Fig. 2.6b). In two subsequent analyses we applied a skewed Cauchy distribution such that the mode shifted towards the minimum or the maximum constraint, reflecting prior views that the fossils used to calibrate the tree are either very good (Fig. 2.6c) or very poor (Fig. 2.6d) proxies of the true divergence times. Our results proved robust to the use of different calibration strategies, only identifying some variability in the size of the recovered credibility intervals (Fig. 2.7a-c).

We explored the impact of different strategies for modelling both the molecular clock (Fig. 2.6e) and the amino acid substitution process (Fig. 2.6f) Only minimal differences in posterior ages were found between analyses using an uncorrelated or autocorrelated clock (Fig. 2.7d) Consistently, Bayesian cross-validation indicated that the two models do not differ significantly in their fit to the data (cross-validation score = 0.7 ± 2.96816 in favour of the uncorrelated clock). In contrast, using a single substitution model across the 29 genes or using an optimal set of gene-specific substitution models inferred using PartitionFinder (Lanfear et al. 2012) resulted in very different age estimates (Figs. 2.6f and 2.7e). Using a single substitution model recovered larger credibility intervals (Fig. 2.6e) with a more homogeneous distribution of branch lengths across the tree, and older divergence times (compare Fig. 2.6f and Fig. 2.7a-d). An Akaike information criterion (AIC) test indicated that the partitioned model provides a significantly better fit to the data (AIC score = 565.21 in favour of 29 gene-specific models), allowing the rejection of the divergence times obtained with a single substitution model. As expected, divergence times estimated from individual genes were much less precise, although posterior age estimates overlap well (Fig. 2.8).

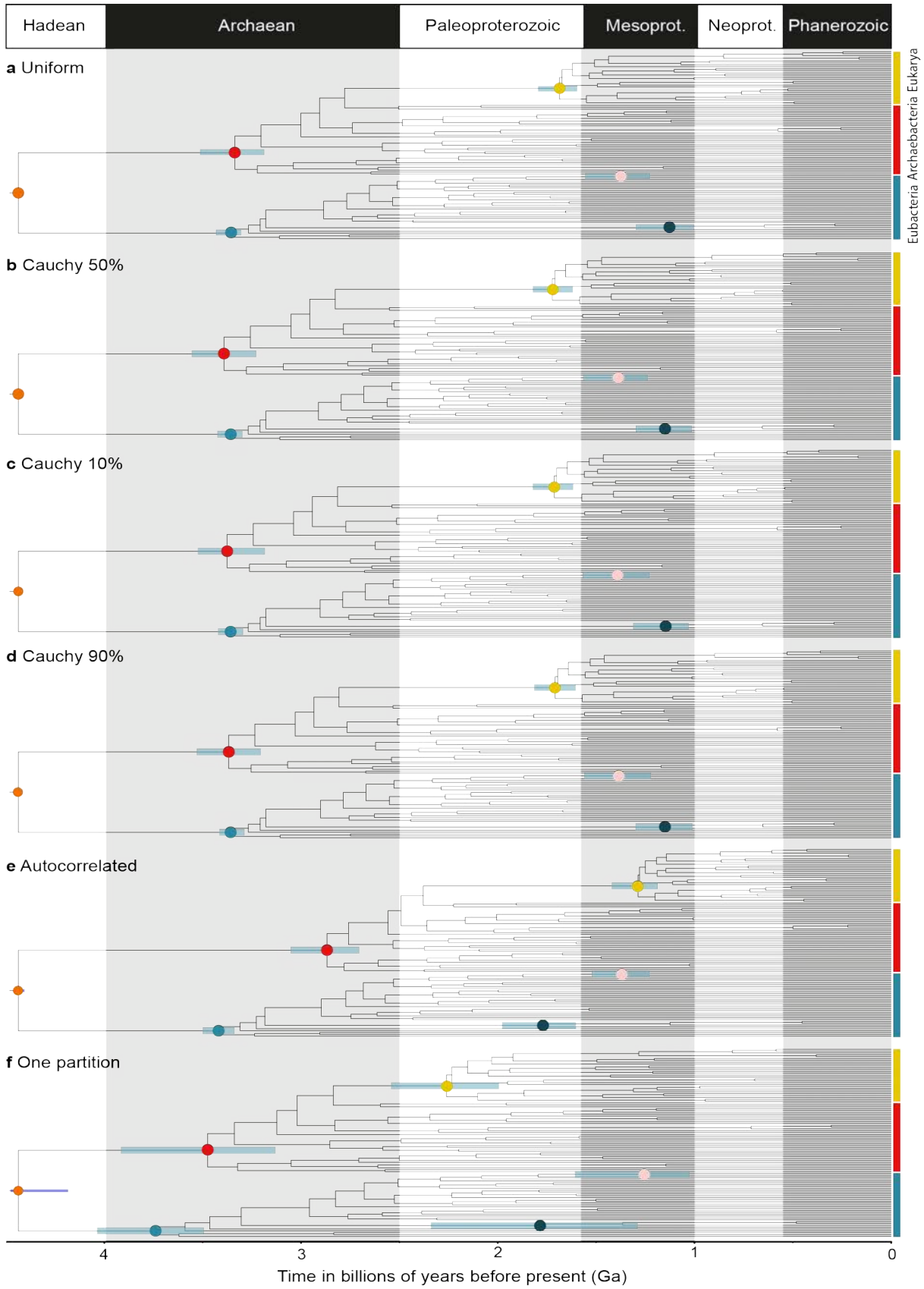


Figure 2.6. Posterior time estimates under different parameters. a, Posterior time estimates when using a uniform calibration density prior distribution, reflecting a lack of information about the divergence time relative to the fossil constraint. b, Cauchy 50% maximum calibration density prior distribution, reflecting a view that the

divergence date should fall between the constraints. c, Cauchy 10% maximum calibration density prior distribution, reflecting a view that the fossil prior is a good approximation of the divergence date. d, Cauchy 90% maximum calibration density prior distribution, reflecting a view that the fossil prior is a poor approximation of the divergence date, all with an uncorrelated clock model. e,f, Posterior age estimates when using a Cauchy 50% maximum calibration density prior distribution with an autocorrelated clock model (e) and with an uncorrelated clock model and a single partition scheme (f). All molecular clock analyses converged well. The coloured dots highlight specific nodes, with their respective confidence intervals displayed light blue bars (orange, LUCA; red, crown Archaebacteria; blue, crown Eubacteria; yellow, crown Eukaryota; pink, Alphaproteobacteria; dark blue, Cyanobacteria). This figure illustrates how divergence times change as alternative approaches to modelling calibrations and the process of molecular evolution are implemented. Divergence estimates from f and their credibility intervals could be rejected based on an AIC test. The other results (a–e) cannot be rejected. Mesoprot., Mezoproterozoic; Neoprot., Neoproterozoic.

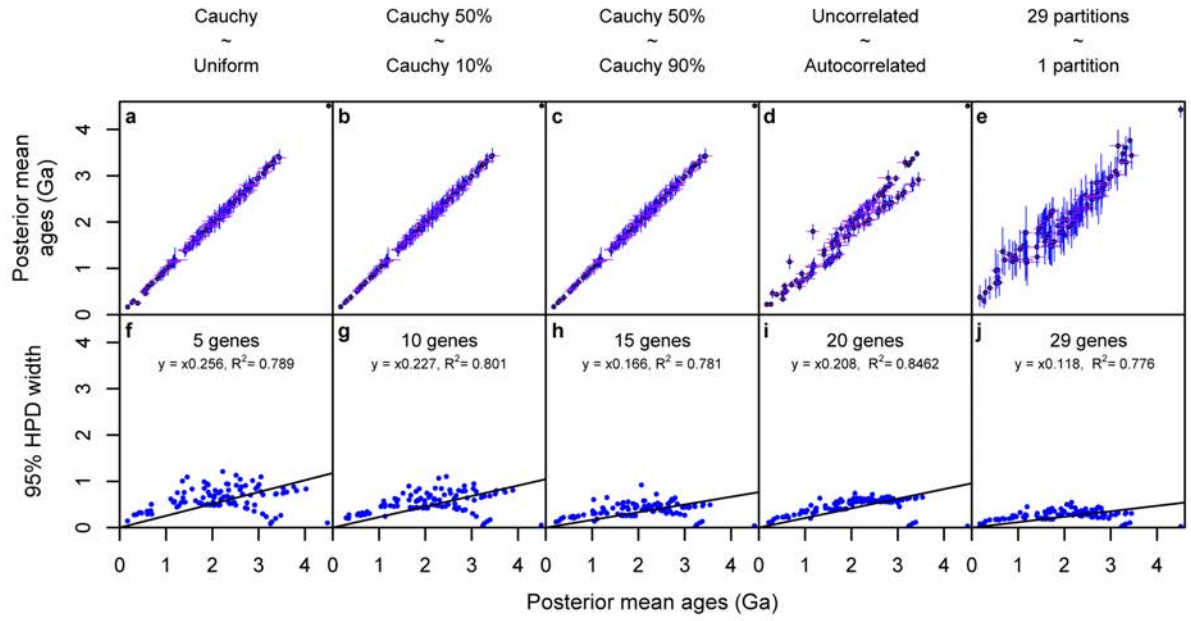


Figure 2.7. Changes in divergence times (Ga) that result from applying alternative parameters. a, Cauchy 50% maximum calibration density prior distribution versus uniform calibration density prior distribution. b, Cauchy 50% maximum calibration density prior distribution versus Cauchy 10% maximum calibration density prior distribution. c, Cauchy 50% maximum calibration density prior distribution versus Cauchy 90% maximum calibration density prior distribution. d, Cauchy 50% maximum calibration density prior distribution uncorrelated clock model versus Cauchy 50% maximum calibration density prior distribution autocorrelated clock model. e, Cauchy 50% maximum calibration density prior distribution in both cases for the 29-partition scheme versus the 1-partition scheme. f–j, Results of adding additional genes as infinite sites plots: 5-gene dataset (f); 10-gene dataset (g); 15-gene dataset (h); 20-gene dataset (i); 29-gene dataset (j). Blue dots denote node dates. HPD, highest posterior density.

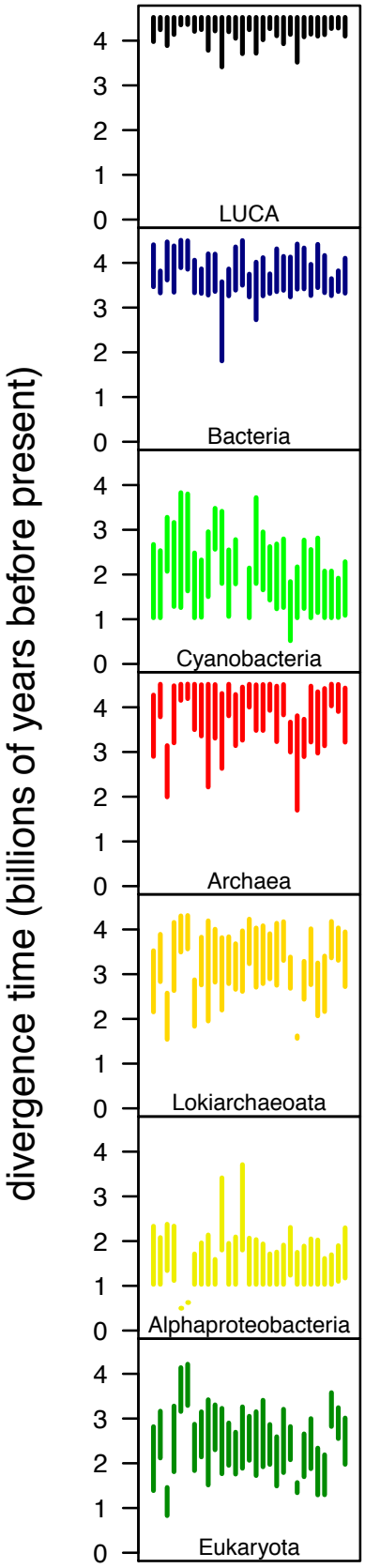


Figure 2.8. Divergence dates for 7 key nodes in the tree of life produced by implementing the molecular clock on a gene by gene basis. In each case a Cauchy 50% calibration distribution density and an uncorrelated clock model was used. On each of the plots the bars represent the divergence dates for genes 1-29.

This indicates that the genes comprising our dataset encode a congruent signal and the timescale inferred from the combined analysis is not biased by single gene outliers. Furthermore, their combination improves the precision of the clade age estimates (Fig. 2.7f–j), which are clearly informed by the data (Fig. 2.9). We tested the effect of taxonomic sampling by doubling the number of Cyanobacteria and Alphaproteobacteria in our dataset. We then explored the effect of phylogenetic uncertainty by dating a tree compatible with Woese's three-domains hypothesis (Woese and Fox 1977) and by dating all 15 trees in the 95% credible set of trees from our phylogenetic analysis (Fig. 2.10 and 2.11). Further analyses that used co-estimation of tree and topology (Figure 2.12) (Drummond et al. 2006) did not reach convergence (Figure 2.13), but the results recovered were congruent with those obtained from well-converged analyses (Figure 2.11) where topology and time were inferred sequentially. Our analysis which attempted to co-estimate time and topology did not converge and its results, that are consistent with those of Figure 2.12, are invalid. It is unsurprising that this analysis did not converge considering the limited amount of information available to date the entirety of the tree of life and the inherent, significant, increased complexity associated with attempting to concomitantly estimate, from these data both a phylogeny and its associated divergence times. The similarity between the results of the well converged analyses reported in Figure 2.11 (that used the trees in the 95% credibility set from our phylogenetic analysis) and the results of the co-estimation analysis Figure 2.13 raise some doubt on the general utility of co-estimation analysis that does not seem to highlight anything than significantly less computationally intense, multiple serial divergence time analyses (using trees from the 95% credibility set) cannot highlight.). Overall, the outcome of these experiments demonstrates that our original results are robust to topological uncertainty and the use of differential taxonomic sampling (Fig. 2.10-2.13).

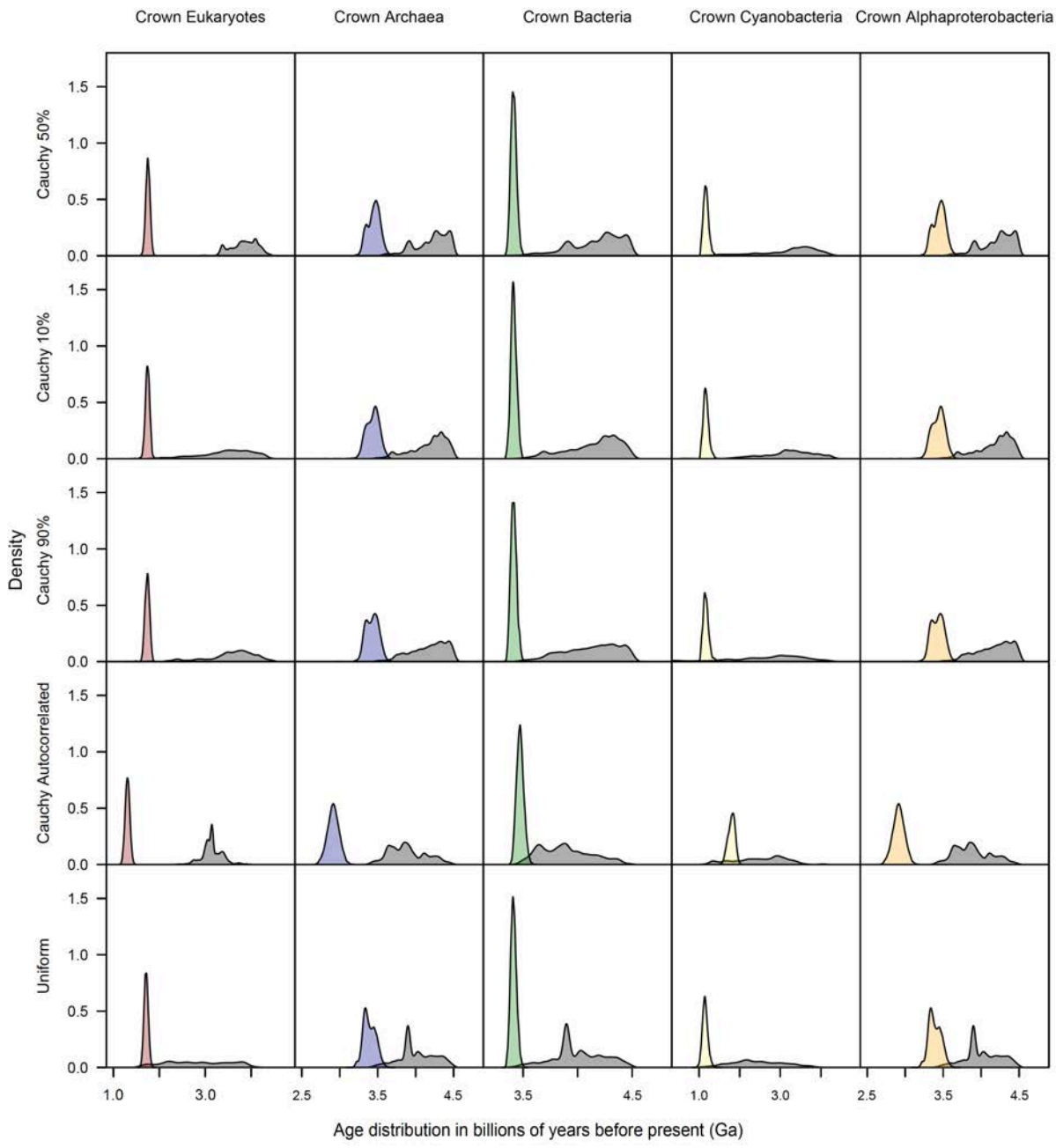


Figure 2.9. Density plots comparing the prior (grey) and the posterior distributions (colour) in divergence times for 5 nodes in the tree of life. The different calibration density distributions and clock models used are listed along the right side.

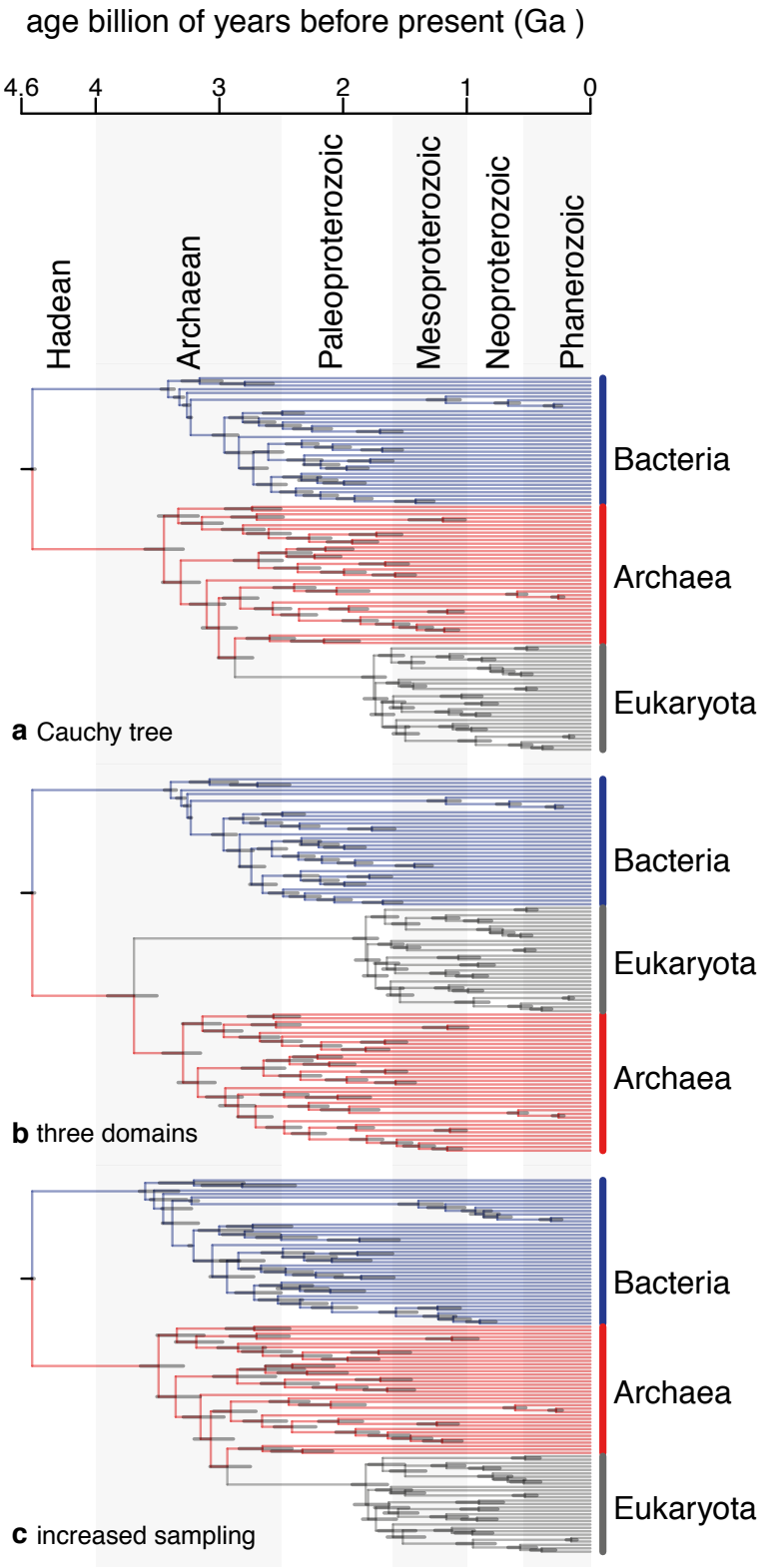


Figure 2.10. Comparison of divergence dates produced using (a) a Cauchy 50% calibration distribution density with Eocyte topology (see also Figure 2.6a), (b) a Cauchy 50% calibration distribution density with a Three Domain Topology, and (c) a Cauchy 50% calibration distribution density with additional species in Alphaproteobacteria and Cyanobacteria. The Eukaryota are highlighted in grey, the Archaebacteria in red and the Eubacteria in blue.

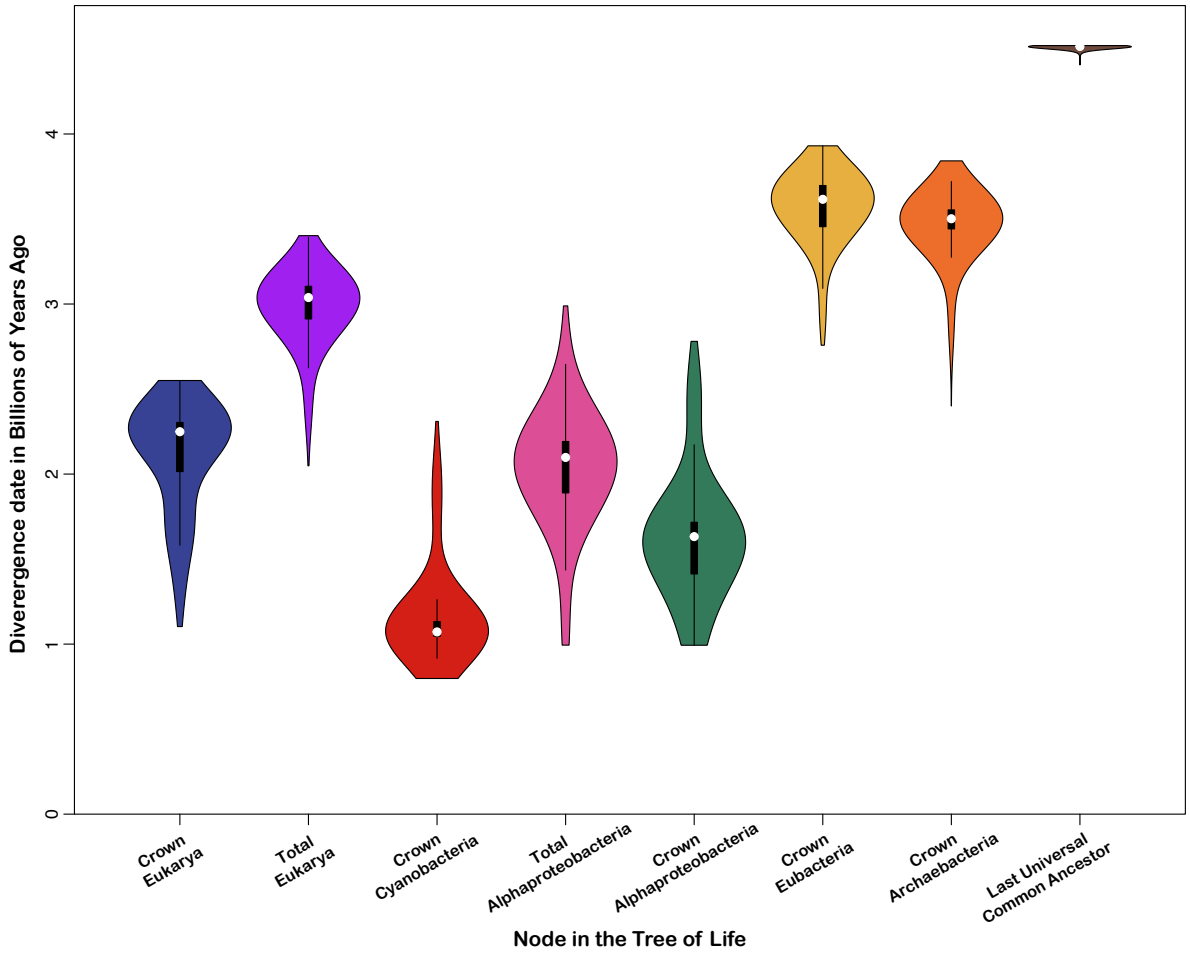


Figure 2.11. Violin plots showing the spread of divergence dates for key nodes in the tree of life from 20 different analyses: Cauchy 50% calibration distribution density; Cauchy 10% calibration distribution density; Cauchy 90% calibration distribution density; Cauchy 50% calibration distribution density with an autocorrelated clock model; Uniform calibration distribution density; and the 15 tree topologies in the 95% credible set of trees from our original PhyloBayes analysis.

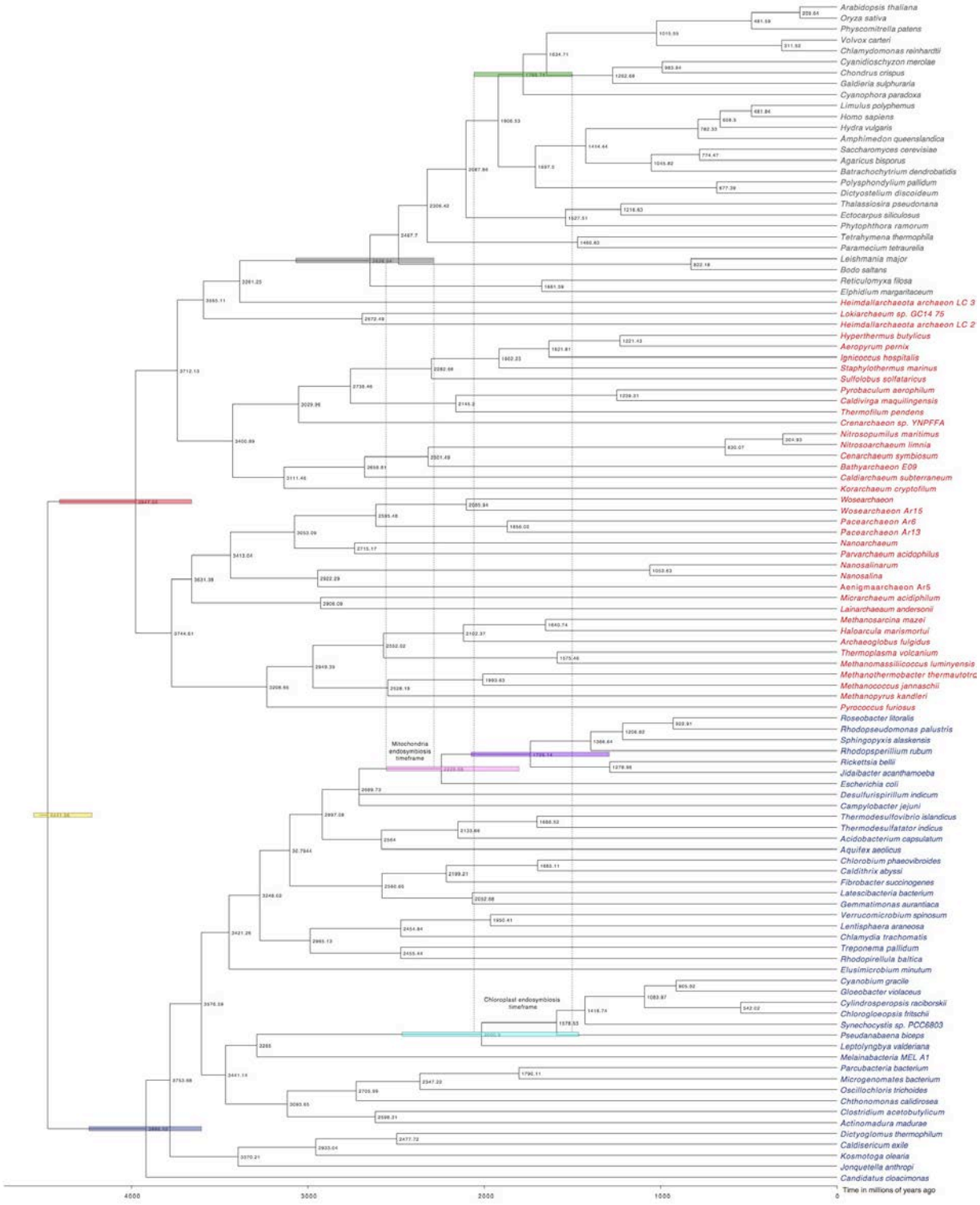


Figure 2.12. Results obtained from an attempt at co-estimating time and topology (20,000,000 generations). The coloured node bars indicate nodes of interest and their 95% posterior credibility interval; green (Archaeplastida), grey (crown Eukaryota), red (crown Archaebacteria), yellow (LUCA), dark blue (crown Eubacteria), pale blue (Cyanobacteria), pink (total Alphaproteobacteria), and purple (crown Alphaproteobacteria).

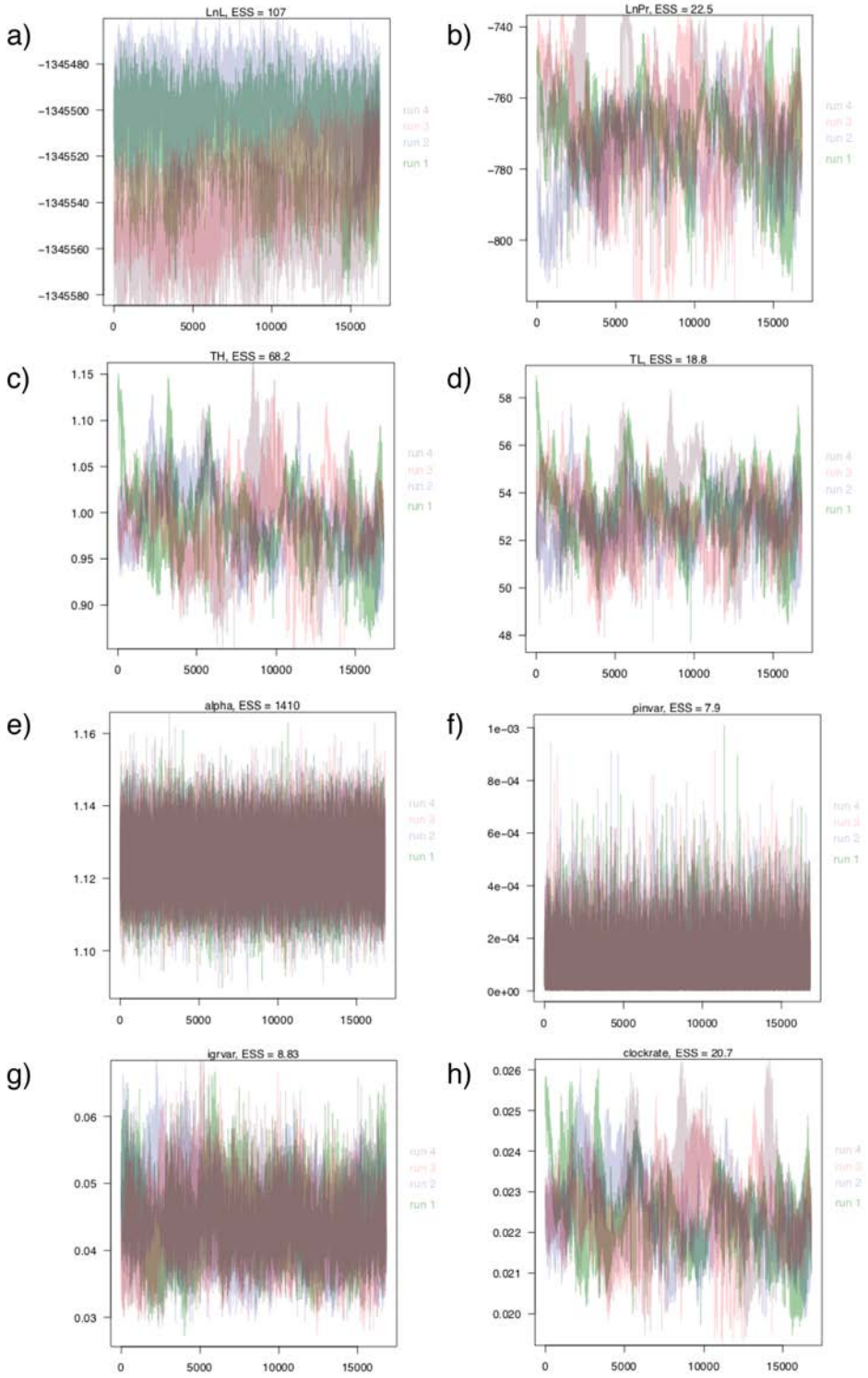


Figure 2.13. Convergence statistics for the co-estimation of time and topology analyses. Traces and ESS (effective sample size) scores (after 20,000,000) clearly indicate that the analysis is still far from convergence. In each case the X-axis is the number of generations and the Y-axis is the parameter values. The different colours represent the 4 different runs. The plots detail the following parameters; a) LnL = log likelihood, b) LnPr = log prior probability, c) TH = tree height, d) TL = tree length, e) alpha = shape parameter of the gamma distribution governing rates across sites, f) pinvar = proportion of invariable sites, g) igvar = Variance increase of igr model branch lengths, and h) clockrate = base rate of clock

2.4 Discussion

It is not possible to discriminate between the competing calibration strategies that reflect different interpretations of the fossil record. Similarly, our model selection test indicated that the auto- correlated and independent-rates clock models fit the data equally well. Thus, in establishing an accurate timescale of life, we integrated over the uncertainties associated with the results from all these analyses (Fig. 2.14). The joint 95% credibility intervals reject a post- late heavy bombardment (~3,900 million years ago (Ma) (Chapman, Cohen, and Grinspoon 2007) emergence of LUCA (4,519–4,477 Ma). The crown clades of the primary divisions of life, Archaebacteria and Eubacteria emerged over one billion years after LUCA in the Mesoarchaean–Neoarchaean. The earliest conclusive evidence of cellular life (Strelley Pool Formation, Australia (Sugitani, Mimura, Takeuchi, Lepot, et al. 2015) falls within the 95% credibility intervals for the ages of the last common ancestors of both clades, indicating that these fossils might belong to one of the two living prokaryotic lineages.

Methanogenesis is classically associated with Euryarchaeota. Our estimate for the age of crown Euryarchaeota (2,881–2,425 Ma) is consistent with carbon isotope excursions indicating the presence of methanogens by 2 Ga (Hayes 1994), but is substantially younger than the earliest possible evidence of biogenic methane in the geochemical record at ~3.5 Ga (Ueno et al. 2006; Wolfe and Fournier 2018). If the geochemical evidence is correct, our timescale implies that methanogenesis predated the origin of Euryarchaeota. This hypothesis would be consistent with recent environmental genomic surveys indicating that other archaeal lineages may also be capable of methane metabolism (Evans et al. 2015) or methanogenesis (Vanwonterghem et al. 2016), and that metabolisms using the Wood–Ljungdahl pathway to fix carbon minimally evolved in stem archaebacteria (Weiss et al. 2016; Williams et al. 2017) and might have been a characteristic of LUCA (Borrel, Adam, and Gribaldo 2016; Sousa, Nelson-Sathi, and Martin 2016; Weiss et al. 2016).

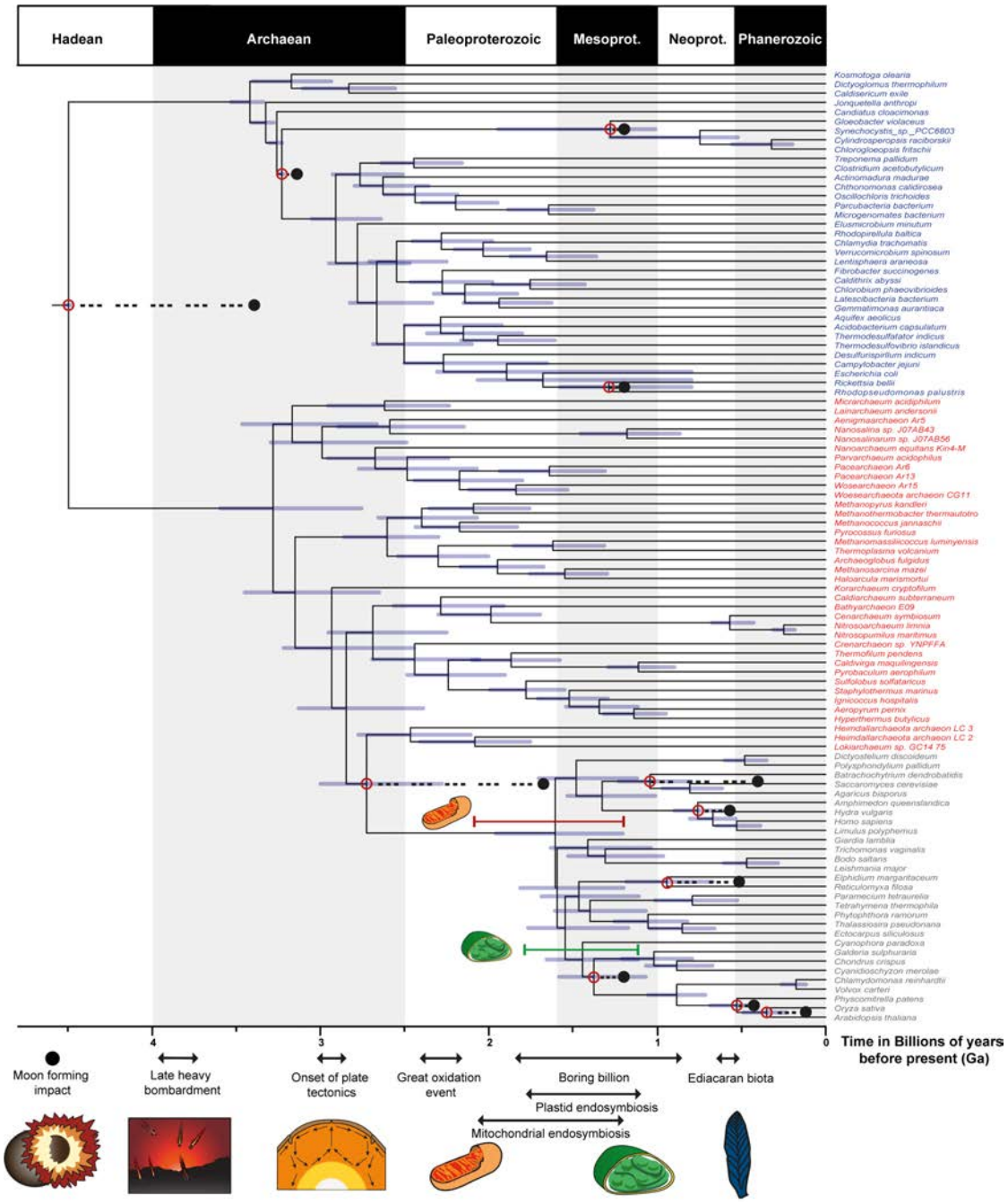


Figure 2.14. A divergence time tree combining uncertainties from approaches using uncorrelated and autocorrelated clock models and different calibration density distributions. Tip labels are shown for Eukaryota (grey), Archaebacteria (red) and Eubacteria (blue). The purple bars denote the credible intervals for each node. Red dots highlight calibrated nodes, and corresponding black dots highlight the age of the minimum bound of its corresponding calibration. The phylogenetic relationships of the mitochondrion within Alphaproteobacteria are still debated (Esser et al. 2004; Fitzpatrick, Creevey, and McInerney 2005; Martijn et al. 2018; Roger, Muñoz-Gómez, and Kamikawa 2017), and it is unclear whether the free-living ancestor of the mitochondrion was a crown or stem representative of this group. The red bar above the crown eukaryote node denotes the time period during which the mitochondrial endosymbiosis may have occurred. The green bar denotes the time during which the plastid endosymbiosis may have occurred. Important events in Earth and life history are indicated along the base of the figure. Mesoprot., Mezoproterozoic; Neoprot., Neoproterozoic.

The Great Oxidation Event (GOE; ~2.4 Ga) was perhaps the most significant episode in the Proterozoic (Lyons, Reinhard, and Planavsky 2014), fundamentally changing the chemistry of Earth's atmosphere and oceans, and probably altering temperature. It has been causally associated with the evolution of Cyanobacteria, as a consequence of their oxygen release (Sánchez-Baracaldo et al. 2017; Schirrmeister et al. 2013) and implicated as an extrinsic driver of eukaryotic evolution (Knoll and Nowak 2017). Our timescale indicates that crown Cyanobacteria and crown Eukaryota significantly postdate the GOE. Crown Cyanobacteria diverged 1,947–1,023 Ma, precluding the possibility that oxygenic photosynthesis emerged in the cyanobacterial crown ancestor. However, the Cyanobacteria separated from other eubacterial lineages (Fig. 2.3), including the non-photosynthetic sister group of the Cyanobacteria (Melainabacteria; Figure 2.10c) in the Archaean, before the GOE, consistent with the view that oxygenic photosynthesis evolved along the cyanobacterial stem (Shih and Matzke 2013), and compatible with a causal role of the total-group Cyanobacteria in the GOE. Crown Eukaryota diverged considerably after both the Eukaryota–Asgardarchaeota split and the GOE, in the middle Proterozoic (1,842–1,210 Ma). Our study strongly rejects the idea that eukaryotes might be as old as, or older than, prokaryotes (Kurland, Collins, and Penny 2006), and agrees with a number of other studies that date the last eukaryote common ancestor (LECA) to the Proterozoic (~1,866–1,679 Ma) (Chernikova et al. 2011; Eme et al. 2014; Parfrey et al. 2011). Within eukaryotes, the main extant clades emerged by the middle Proterozoic, including Opisthokonta (~1,707–1,125 Ma), Archaeplastida (~1,667–1,118 Ma) and SAR (stramenopiles (heterokonts), alveolates and foraminifera; ~1,645–1,115 Ma). The symbiotic origin of the plastid occurred among stem archaeplastids (~1,774–1,118 Ma), and our 95% credibility interval for the origin of the plastid overlap with the results of other recent studies (Sánchez-Baracaldo et al. 2017; Shih and Matzke 2013). The relatively long stem lineage subtending LECA is intriguing. It is found using both uncorrelated and autocorrelated clock models (Figs. 2.6e and 2.7d) and, disappears only if a poorly fitting single substitution model is used (Figs. 2.6f and 2.7e), suggesting that it is not a modelling artefact. Analyses excluding the hitherto unknown immediate living relatives of Eukaryota (Spang et al. 2015; Zaremba-Niedzwiedzka et al. 2017), Asgardarchaeota, had no significant impact on the span of the eukaryote stem lineage, suggesting that its length is robust to taxon sampling (Fig. 2.15).

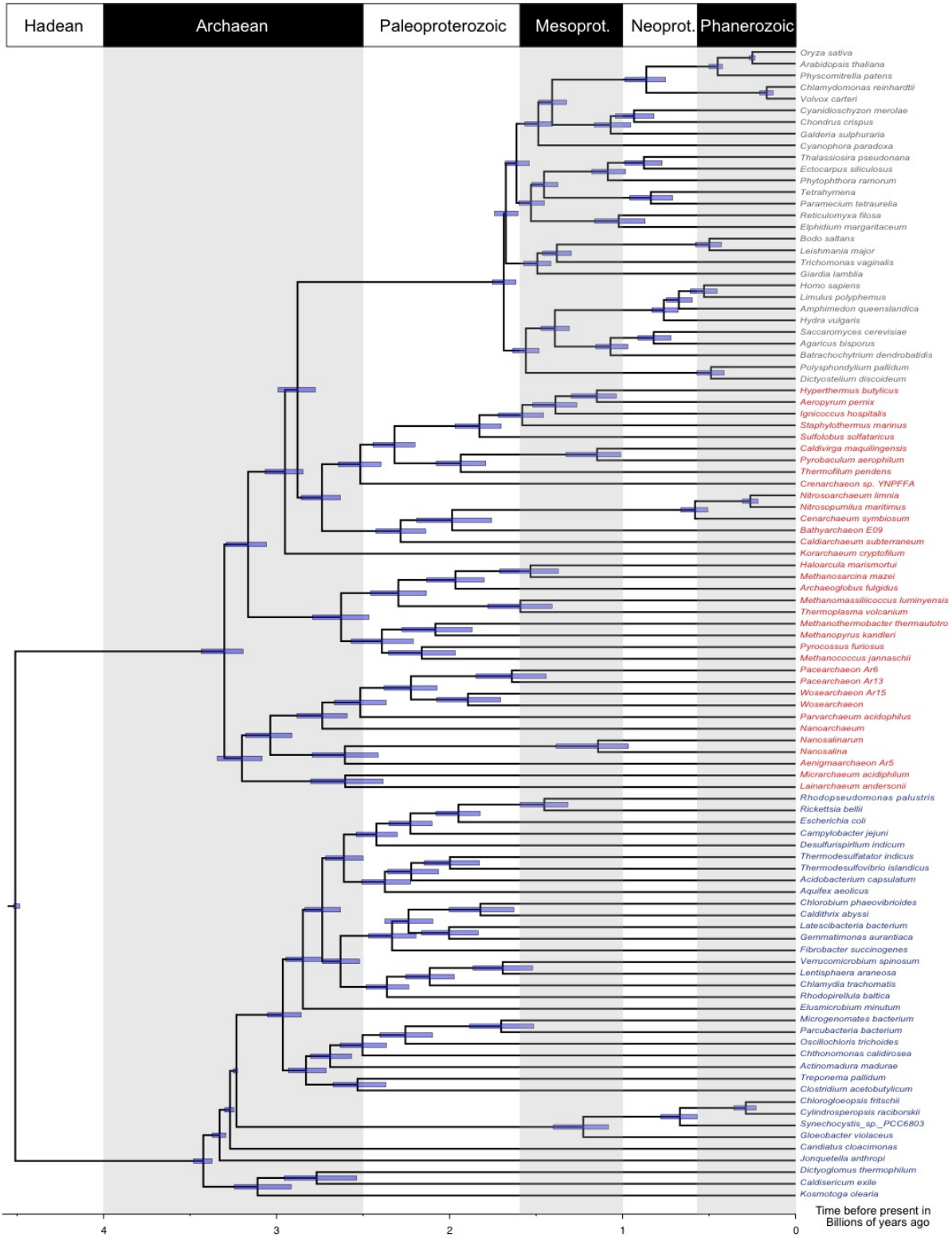


Figure 2.15. Divergence times produced using a Cauchy 50% calibration density distribution and an uncorrelated clock model with the Asgardarchaeota removed. The Eukaryota are highlighted in grey, the Archaebacteria in red and the Eubacteria in blue.

Our timescale for eukaryogenesis rejects the hypothesis of an inextricable link between the GOE and the origin of eukaryotes (Knoll and Nowak 2017). Competing hypotheses for eukaryogenesis hinge on the early versus late acquisition of mitochondria relative to other key eukaryote characters (Martin et al. 2017; McInerney, O'Connell, and Pisani 2014; Pittis and Gabaldón 2016; Roger, Muñoz-Gómez, and Kamikawa 2017; Pisani, Cotton, and McInerney 2007). Absolute divergence times cannot discriminate between these hypotheses. However, as the only proposed evidence in support of the mitochondria late (Pittis and Gabaldón 2016) hypothesis have been argued to be artefactual (Martin et al. 2017), the similar age estimates for Alphaproteobacteria and LECA at this stage are most conservatively interpreted as indicating that the process of mitochondrial symbiosis underpinned a rapid process of eukaryogenesis. This process involved a large transfer of genes from the genome of the alphaproteobacterial symbiont to that of the archaeal host (Ku et al. 2015; Pisani, Cotton, and McInerney 2007), as predicated on metabolism (McInerney, O'Connell, and Pisani 2014; Lane and Martin 2010).The search for the earliest fossil evidence of life on Earth has created more heat than light. Although the fossil record remains integral to establishing a timescale for the Tree of Life, it is not sufficient in and of itself. Our integrative molecular timescale encompasses the uncertainty associated with fossil, geological and molecular evidence, as well its modelling, allowing it to serve as a solid foundation for testing evolutionary hypotheses in deep time for clades that do not have a credible fossil record.

Chapter 3

The application of cross-bracing using ancient gene duplications to date a tree of life

Author contributions: The ideas for this chapter were developed by D. Pisani, P.C.J. Donoghue, T.A. Williams and H.C. Betts. The dataset was collected, and all analyses carried out by H.C.B. with help and suggestions from D.P., P.C.J.D. and T.A.W. A new version of PAML (v 4.9i) was provided by Z. Yang to test in this study. H.C.B. wrote and developed the following chapter with comments and suggestions from the other authors. H.C.B. contributed to 90% of the work in this Chapter.

3.1 Introduction

In Chapter 2 of this thesis I outlined an integrative approach to dating the tree of life which employed a dataset of 29 concatenated genes, 103 species, and 11 fossil calibrations in a node calibration framework. While this approach produced a robust timescale and allowed for the assessment of a diversity of evolutionary hypotheses, it lacked power to accurately infer the age of the oldest node, the last universal common ancestor (LUCA). This is because there is no outgroup for life and so no information for the branch leading to LUCA (Fig. 3.1a). Indeed, the age of the root in any dated phylogeny is invariably the most likely to be incorrect, one of the reasons why a prior on the root age is necessary in Bayesian molecular clock analyses (Bouckaert et al. 2014; Lartillot, Lepage, and Blanquart 2009; Yang 2007). In addition, in comparison to most molecular clock analyses which focus on younger lineages, the lack of information for the branch leading to LUCA is likely to be exacerbated due to the evolutionary disparity between the two daughter lineages, Archaebacteria, and Eubacteria. For the deepest node in the tree of life, information contained in the sequence data is poor, almost by definition.

In order to try and overcome this problem we used gene trees which possess a probable pre-LUCA duplication, both singly and as a concatenated dataset, as a means to investigate the age of LUCA with more clarity. Similarities between certain groups of proteins across all life have been recognised for a while, setting the stage for the investigation of ancient duplication events (Schwartz and Dayhoff 1978). Investigating the root of the tree of life is problematic, but an exception to this can be made when a gene duplication produces sequences ancestral to the node of interest. Thus, the gene trees for the two paralogs can be reciprocally rooted using the other paralog group as an outgroup (Fig. 3.1b). This approach was used in the very first studies that attempted to root the tree of life and in a series of subsequent studies aiming to do the same thing (Gogarten et al. 1989; Philippe and Forterre 1999; Forterre and Philippe 1999; Zhaxybayeva, Lapierre, and Gogarten 2005; Iwabe et al. 1989; Lopez, Forterre, and Philippe 1999; Lawson, Charlebois, and Dillon 1996; Charlebois et al. 1997; Gribaldo and Cammarano 1998; Labedan et al. 1999; Brown and Doolittle 1995; Brown et al. 1997).

Genes with proposed pre-LUCA duplications often have important functions within the cell, such as the V- and F-type ATPases subunits (Gogarten et al. 1989), translation elongation factors EF-Tu/1 and EF-G/2 (Iwabe et al. 1989) and a couple of groupings of amino-acyl tRNA synthetases, valyl- and leucyl-tRNA synthetases and tryptophanyl- and tyrosyl- tRNA synthetases (Brown and Doolittle 1995; Brown et al. 1997). While these studies have been inconclusive about the root of life, placing it most consistently on a branch leading to the Eubacteria (Brown and Doolittle 1995; Brown et al. 1997; Gogarten et al. 1989; Gribaldo and Cammarano 1998; Iwabe et al. 1989; Labedan et al. 1999; Lawson, Charlebois, and Dillon 1996) but sometimes in another position (Charlebois et al. 1997; Forterre and Philippe 1999; Lopez, Forterre, and Philippe 1999; Philippe and Forterre 1999), they do all demonstrate that the genes are present in each of the major lineages. In some cases, there may even be more than one duplication event leading prior to LUCA such as can be viewed in theory in Figure 3.1c.

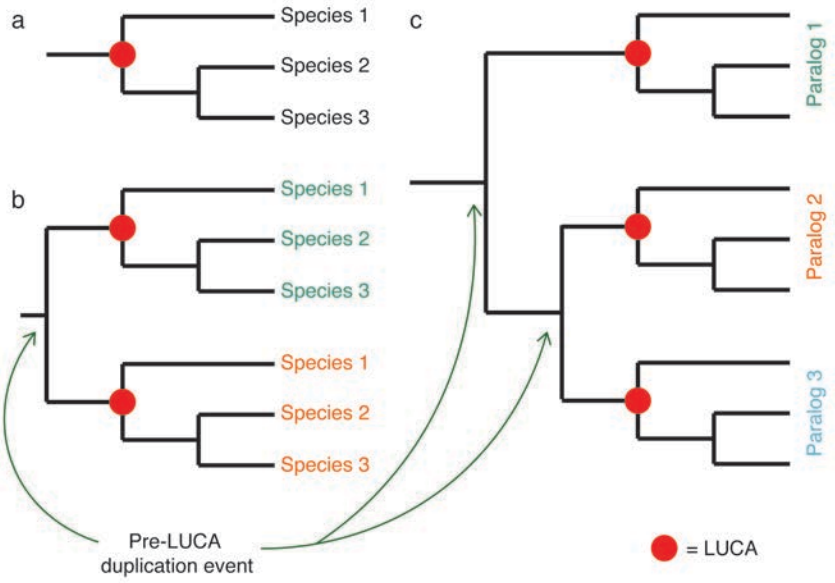


Figure 3.1. Illustration of a duplicated node leading to two or more paralogs. a) A gene tree with no duplications, the root can either be arbitrarily decided or through the use of software, b) a single duplication resulting in two paralagous genes. This means that there is one pre-LUCA node and c) two duplication events prior to LUCA meaning that there are 3 paralagous genes.

In addition to removing LUCA from the root node, families with duplications that precede LUCA are attractive because they permit the placement of the same calibration on either side, allowing for much better constraints on estimation of rates of evolution (Shih and Matzke 2013). The two calibration points for each speciation event provide an extra source of material for dating the tree. There are two ways in which corresponding nodes on both sides of a phylogeny can be calibrated as introduced by Shih and Matzke in 2013. The first is cross calibration, where each node is given the same prior information, but the posterior distribution produced is allowed to vary. The second is cross-bracing. In this approach for every speciation event replicated on either side of the tree the same calibration can be applied and 'braced' such that it must have the exact same age on both sides of the tree. This is biologically realistic as we expect speciation events to have occurred at the same time. Cross-calibration has been used to date whole genome duplications in plants (Clark and Donoghue 2017) as well as the timing of endosymbiosis events using the V- and F-type ATPases (Shih and Matzke 2013). The latter have also been dated using cross-bracing in the same study. Here, we use cross-bracing on ancient gene duplications, alongside a set of robust fossil calibrations, to gain more power to estimate a timescale for the tree of life, and, in particular, the last universal common ancestor.

3.2 Materials and Methods

3.2.1 Dataset collection and phylogenetic analyses

The 8 gene families investigated in this study are listed in Table 3.1. These have all previously been used to investigate the root of the tree of life and have been shown likely to possess a duplication prior to LUCA (Brown and Doolittle 1995; Brown et al. 1997; Gogarten et al. 1989; Forterre and Philippe 1999; Philippe and Forterre 1999; Zhaxybayeva, Lapierre, and Gogarten 2005; Iwabe et al. 1989; Lopez, Forterre, and Philippe 1999; Lawson, Charlebois, and Dillon 1996; Charlebois et al. 1997; Gribaldo and Cammarano 1998; Labedan et al. 1999). Hence, making them good candidates for this study. For each gene family we extracted data for a chosen set of species from NCBI using BLAST (Altschul et al. 1990). Sequences were selected to create a broad range of taxonomic sampling across the tree of life with a focus on eukaryote species for the purpose of applying calibrations. For each gene family the sequences were aligned using MUSCLE (Edgar 2004) and trimmed using TrimAl (Capella-Gutiérrez, Silla-Martínez, and Gabaldón 2009) using the -strict setting. Gene tree phylogenies were estimated using the maximum likelihood software IQTree (Nguyen et al. 2014). Model finder was used as an IQTree option and the best fitting model (Kalyaanamoorthy et al. 2017) was chosen according to the BiC. This model was used where necessary for subsequent analyses (Table 3.1).

At this stage any sequences considered to be erroneous, those with extremely short or long sequences, extremely long branches, or duplicates at the tip of the tree, were removed. If necessary, further rounds of topology generation and cleaning were carried out. Once cleared of erroneous species, a rogue taxon search was performed using RogueNaRok (Aberer, Krompass, and Stamatakis 2012) and taxa with uncertain positions removed. In two cases (EF-Tu/1 and Tyrosyl-tRNA synthetase) the root of one paralog within the gene was uncertain. This was because the topology produced did not conform to an expected LUCA root. However, Archaebacteria and Eubacteria were mostly grouped and the root was placed in one of these, rather than in between them. In order to investigate whether anything was biasing the root position, such as long branch attraction (LBA), a minimal ancestor deviation (MAD) rooting analysis (Tria, Landan, and Dagan 2017) was performed on the paralog of interest. The MAD software

accepts an unrooted tree or set of trees and then using the concept of operational taxonomic units places the root in all positions across the tree. The root position that minimises the deviation from a midpoint root is assumed to be the root node. The software provided an independent assessment of where the root is likely to be.

A concatenated dataset for topological scrutiny was produced using FASconCATv1.0 (Kück and Meusemann 2010). The sequences were concatenated such that each paralog was considered its own gene and no duplication was retained within the resultant topology. Phylogenies were estimated using both maximum likelihood and Bayesian methods in IQTree (Nguyen et al. 2014) and PhyloBayes (Lartillot, Lepage, and Blanquart 2009). Initially, the concatenated dataset contained 6 of the gene families with both paralogs and 1 gene family with only one paralog. We used this to identify rogue sequences. Subsequently, we removed two more paralogs with uncertain phylogenetic histories. This resulted in a final dataset containing 5 complete gene families and 2 with only one paralog. A similar approach was used for a dataset where the duplication was retained, and the genes concatenated such that it would be preserved. Here, maximum likelihood was used to infer a tree using IQTree (Nguyen et al. 2014) with an LG + C60 model. In both cases the complete gene families were the ATPases, elongation factors, signal recognition proteins, Tryptophanyl and Tyrosyl-tRNA synthetases and the Valy-Leucyl- and Methionyl- tRNA synthetases.

3.2.2 Divergence time analyses

In total we used 11 calibrations. These were applied according to Betts et al., 2018 (and Chapter 2 of this thesis. Section 2.3.1). Calibrations were applied wherever all the requisite species for the calibrations were present even if the gene tree topology for these species did not match that of recognised species trees (this was carried out because gene trees, not having as much information as super alignments, might have greater numbers of stochastic errors). A uniform distribution was used for the calibration density distribution in all cases. The estimation of divergence dates was conducted using the cross-bracing method in MCMCTree (Yang 2007) version 4.9i with the approximated likelihood method. Cross-calibration where used was also carried out using MCMCTree in PAML. The

model used to analyse each gene was selected using the IQTree model finder (Kalyaanamoorthy et al. 2017) selection and in all cases a variant of the LG model was preferred (Table 3.1). Calibrations applied to crown nodes were cross-braced, those applied to stem-nodes where the outgroup was the same in each case were cross-braced, if not they were merely cross-calibrated (Figure 3.2). For an explanation of cross-bracing vs cross-calibration see sections 1.6.2 and 3.1. For example, if in both paralogs the outgroup taxa to Cyanobacteria was *Candidatus Melainabacteria* then the calibration was cross-braced such that the posterior distribution has to be the same. However, if the outgroup was found to be different in each paralog, then the nodes were merely cross-calibrated. Thus each, node could have a separate posterior distribution. If MAD was used to find the root of a paralog, the root found by this analysis was used in the divergence time estimation. For all molecular clock analyses, convergence was tested in Tracer (Rambaut et al. 2018) by comparing plots of estimates from the two independent chains and evaluating whether—for each model parameter and divergence time estimate—the effective sample size was sufficiently large. All reported molecular clock analyses reached excellent levels of convergence.

Table 3.1. A table containing a list of the genes used in this study along with their best fitting amino-acid exchange rate matrices as picked by IQTree model finder.

Gene family	Bayesian information criterion (BIC)	Akaike information criterion (AIC)
F- and V- type ATPases (ATPases)	LG+C50+F+R7	LG+C50+F+R7
Carbamoyl phosphatases (CPS)	LG+C50+F+R4	LG+C50+F+R4
Elongation Factors (EF)	LG+C60+F+R8	LG+C60+F+R8
Histidine biosynthesis subunits A and F (HisAF)	LG+C50+F+R6	LG+C50+F+R6
Aspartate/Ornithine carbamoyltransferases (OTC/ATC)	LG+C40+F+R10	LG+C60+F+R10
Signal Recognition Proteins (SRP)	LG+C60+F+R6	LG+C60+F+R7
Tryptophanyl-tRNA and Tyrosyl-tRNA synthetases (Tyr-Trp)	LG+C60+F+R6	LG+C60+F+R6
Valyl-, Methionyl-, and Leucyl- tRNA synthetases (Val-Leu-Met))	LG+C50+F+R10	LG+C50+F+R10

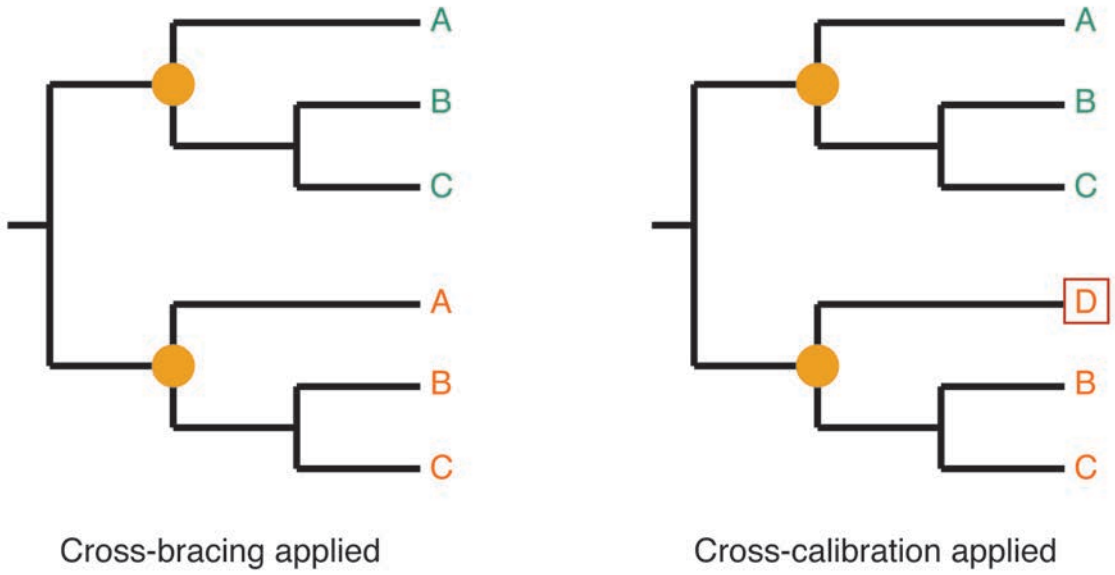


Figure 3.2. Figure illustrating the difference between total-group calibrations when they are cross-braced vs cross-calibrated. In the first tree both of the paralogs have the same species and so the yellow node can be cross-braces. In the second tree the outgroup to species B and C is different between the paralogs. This means that the node is cross-calibrated, not cross-braced. This is the case only when we are applying a calibration to the total-group rather than to the crown-group of a lineage.

3.3 Results

Once initial gene phylogenies had been produced and rounds of cleaning undertaken the final topology of the gene trees were found to vary. In each case slightly different species remain at the tips. Despite this, enough species are retained in each case to be able to investigate the nature of the earliest nodes in the tree. In total 5 of the gene families (F- and V- type ATPases, Elongation Factors , Signal Recognition Proteins, Tryptophanyl-tRNA and Tyrosyl-tRNA synthetases and Valyl-, Methionyl-, and Leucyl-tRNA synthetases) preserve a LUCA root within both paralogs, 2 more (Aspartate/Ornithine carbamoyltransferases and Histidine biosynthesis subunits A and F) have one paralog with a definite LUCA root and 1 (Carbamoyl phosphatases) has no clear LUCA root in either paralog. Though the EF and Tyr-Trp families only produced a LUCA root when analysed using MAD. Here I provide a brief overview of each of the gene families before summarising the results from concatenate analyses.

3.3.1 Topology

3.3.1.1 F- and V- type ATPases

This gene family is found across all three domains of life where the genes help to facilitate active transport across endomembranes. It is divided into catalytic and non-catalytic subunits. The former made up of F-type subunit A plus V-type subunit B and the latter of F-type subunit B and V-type subunit A. The F-type subunits are dominated by Eubacteria and the V-type subunits by Archaebacteria. Throughout the course of refining these genes our analysis found a root between a grouping of F-type alpha and V-type beta, and a grouping of F-type beta and V-type alpha. However, there was also support for a root where the eubacterial sequences (F-type subunits) group together and the archaebacterial sequences (V-type subunits) group together. The latter was found in our final topological analysis although the groupings in this case were very poorly supported (52 Bootstrap probability (BP)) (Fig. 3.3). Thus, when divergence time estimation was carried out, both this root and one where F-type alpha and V-type beta group together and F-type beta and V-type alpha group together were investigated. This means that in the former there is no duplication prior to the LUCA node and instead two duplications afterwards, one prior to Eubacteria and one prior to Archaebacteria. The V-type subunits

are mostly archaebacterial sequences with a few Eubacteria, possibly located there by (lateral gene transfer) LGT events. In both F-type subunit A the chloroplastic sequences emerge within Cyanobacteria as sister to *Pseudanabaena* (Fig. 3.4a) but with low support (55 BP) and in F-type subunit B the chloroplastic sequences also fall within Cyanobacteria (Fig. 3.4b). Likewise, in both subunits the mitochondrial sequences group with Alphaproteobacteria (Fig. 3.4). In V-type subunit B eukaryotes resolve as sister to the 3 included Heimdallarchaeota species with a support of 80 (Fig. 3.3). In the other paralog the eukaryotes position is uncertain grouping next to a mix of archaeal species.

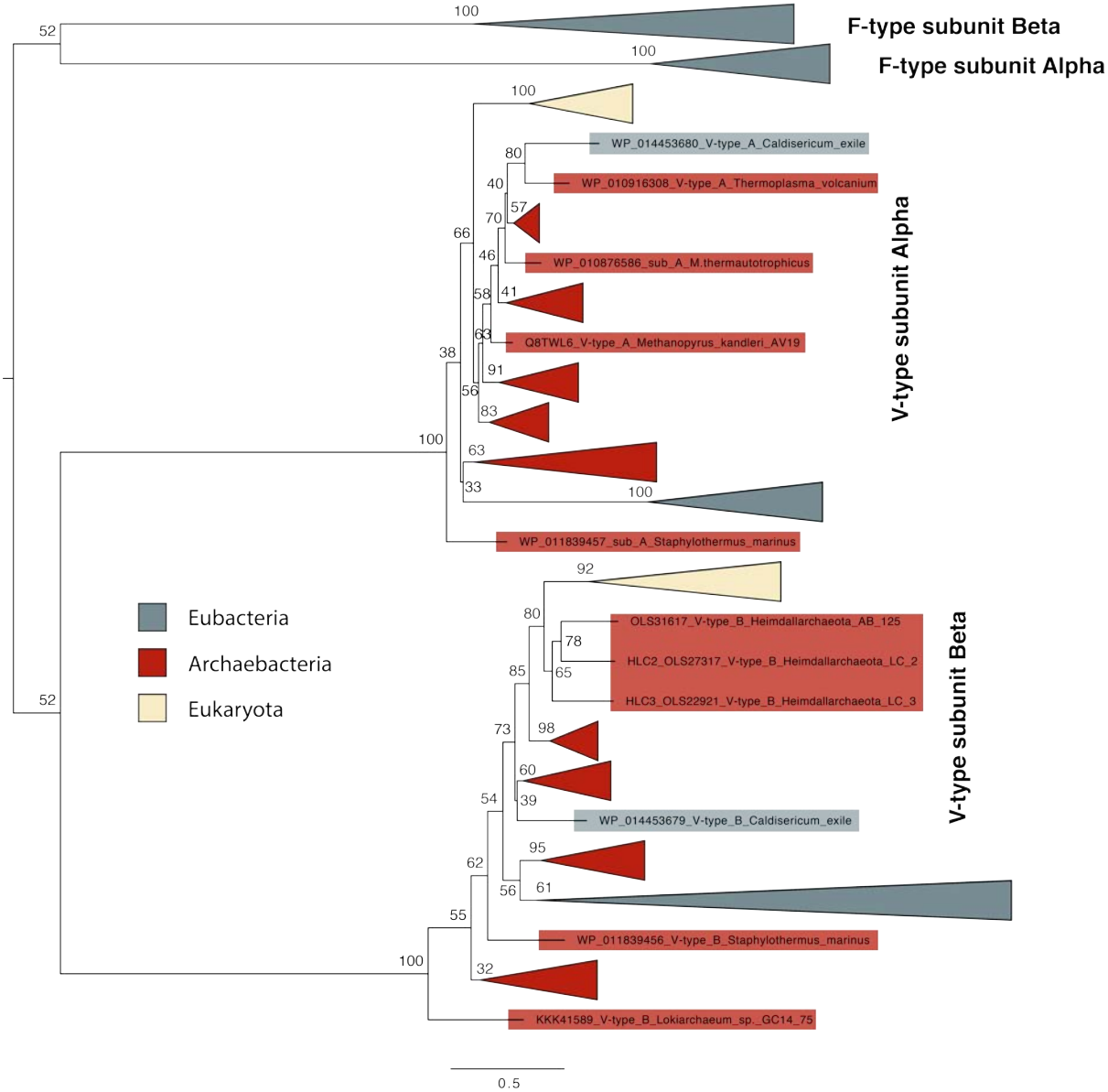


Figure 3.3. Maximum likelihood tree of the F- and V- type ATPases. Support values indicated at the corresponding nodes are the Bootstrap Percentage (BP). Branch lengths are called to the number of character substitutions per site.

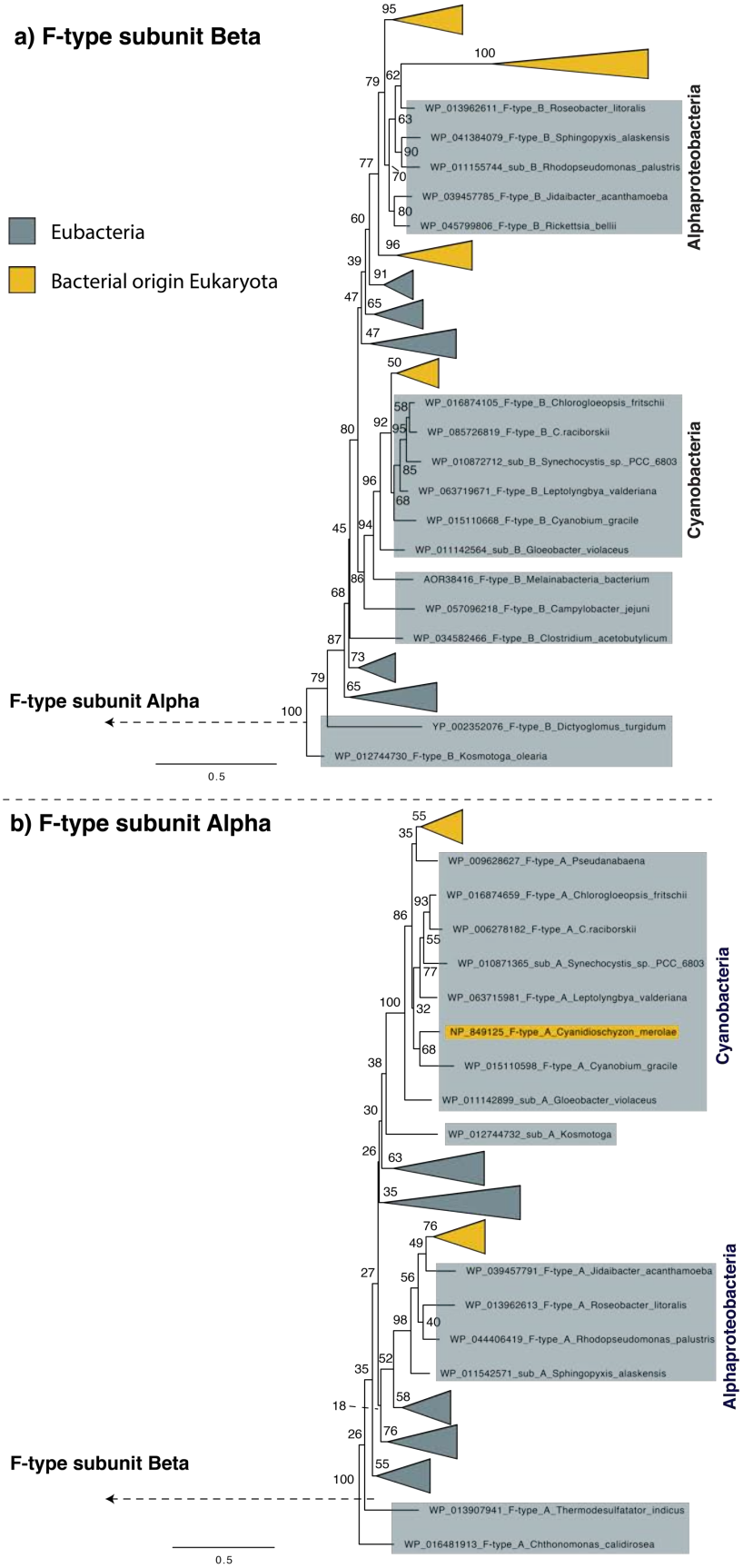


Figure 3.4. Maximum likelihood tree with a focus on the F- type ATPases which contain eubacterial sequences, a) is F-type subunit B and b) F-type subunit A. Support values indicated at the corresponding nodes are the Bootstrap Percentage (BP). Branch lengths are the number of character substitutions per site.

3.3.1.2 Carbamoyl phosphatases

The duplication associated with this gene family is found within the gene and so had to be teased out in a slightly different fashion by cutting each blasted gene at the appropriate sites once aligned. The sites for these cuts were chosen using a previous study (Nyunoya and Lusty 1983). The topology produced had three distinct groupings of taxa, all of which contain some eubacterial, archaebacterial and eukaryotic sequences (Fig. 3.5). However, domain monophyly is not preserved in any case and there is no clear LUCA for any of the sequence groupings. Although there are some amalgamations of eubacterial or archaebacterial sequences, these often have some species from another domain within them. These could be the result of LGTs, and they make it hard to be certain where the root sits in both paralogs. On one side of the duplication in paralog 2 the root falls within Eubacterial sequences and on the other in paralog 1 it falls within archaebacterial sequences. Our results suggest that this gene family should not be used to date the tree of life as there is no clear signal from which to infer the root of the gene family. Hence, we cannot be certain if this gene can be traced back to an ancestor in LUCA. In subsequent analyses we did not attempt to the CPS gene family either to date LUCA individually, or when combined with other genes.

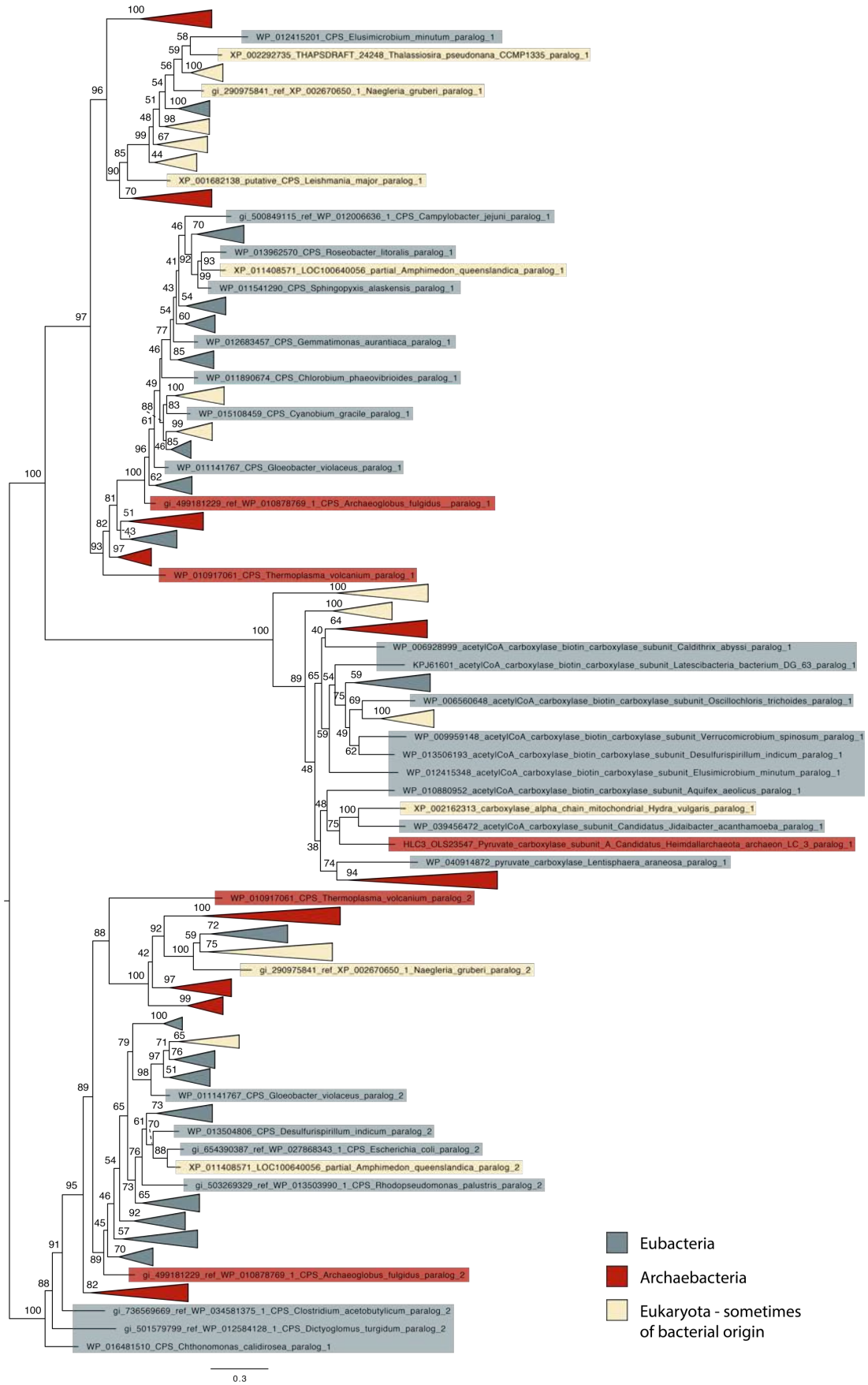


Figure 3.5. Maximum likelihood tree of the carbamoyl phosphate gene family. Support values indicated at the corresponding nodes are the Bootstrap Percentage (BP). Branch lengths are the number of character substitutions per site.

3.3.1.3 Elongation Factors

The two gene families which sit either side of this duplication are EF-Tu/1 and EF-G/2. It is a gene that has been at the centre of controversy regarding the placement of eukaryotes within Archaebacteria (Da Cunha et al. 2017; Spang et al. 2018; Spang et al. 2015). The topology produced is consistent with an Eocyte tree within the EF-G/2 paralog where eukaryotes emerge as sister to Heimdallarchaeota_LC_3, with high support 99 BP, and the root falls between Archaebacteria and Eubacteria (Fig. 3.6). However, within the EF-Tu/1 paralog domain monophyly is not preserved and instead the root falls between the eubacterium *Jonquetella anthropi* and all the other species with high support of 100 BP (Fig. 3.6). To investigate whether this result was real or an artefact resulting from either model inadequacies, or biases created via the long outgroup branches, we ran a maximum likelihood analysis for this gene by itself and with some of the long branches removed. A MAD rooting technique was then applied and a root between Eubacteria and Archaebacteria recovered (Fig. 3.7). Eukaryotes here are found to group as sister to Heimdallarchaeota_AB_125. In divergence time analyses the root found using MAD for EF-Tu/1was used.

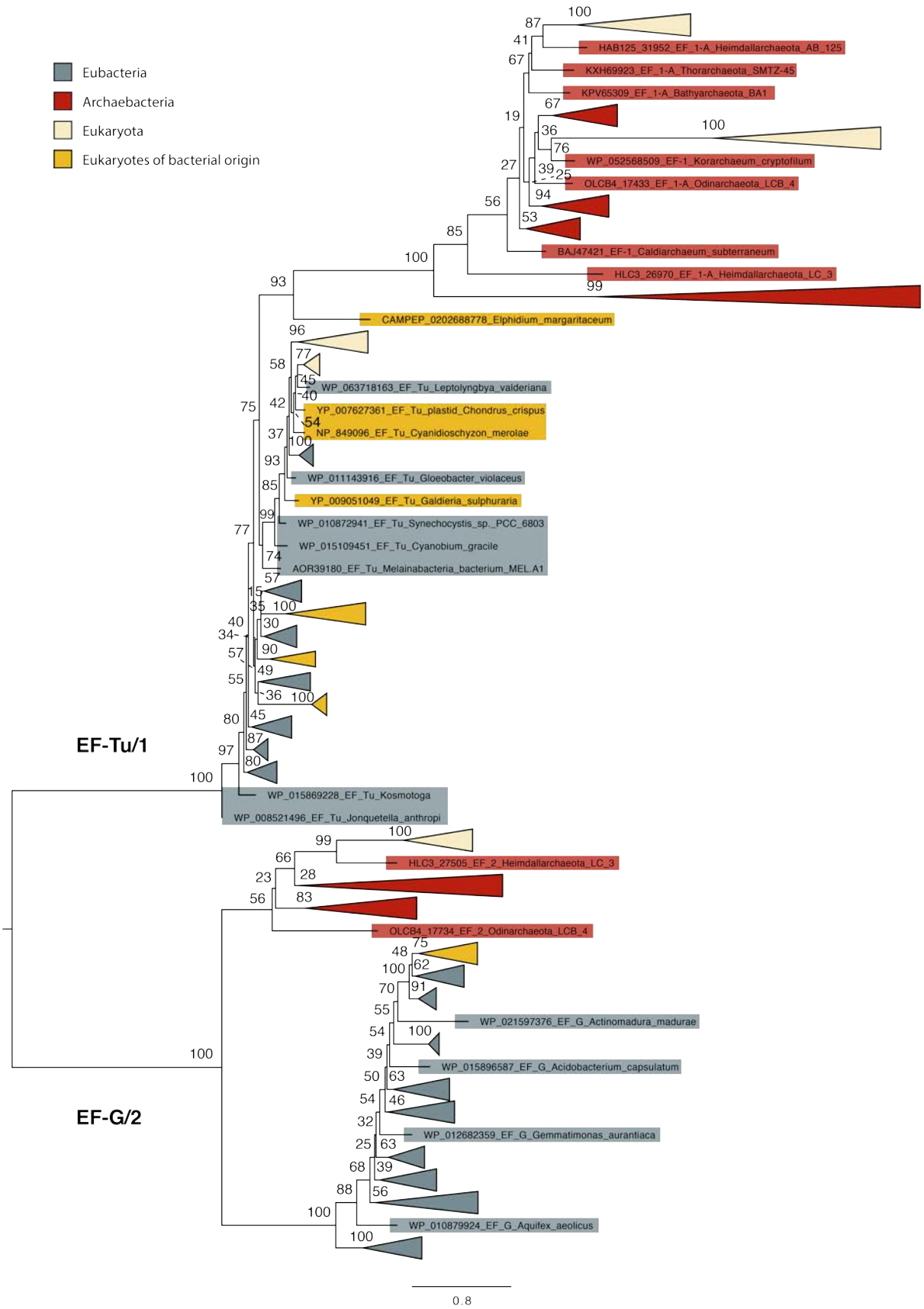


Figure 3.6. Maximum likelihood tree of the elongation factor gene family. Support values indicated at the corresponding nodes are the Bootstrap Percentage (BP). Branch lengths are the number of character substitutions per site.

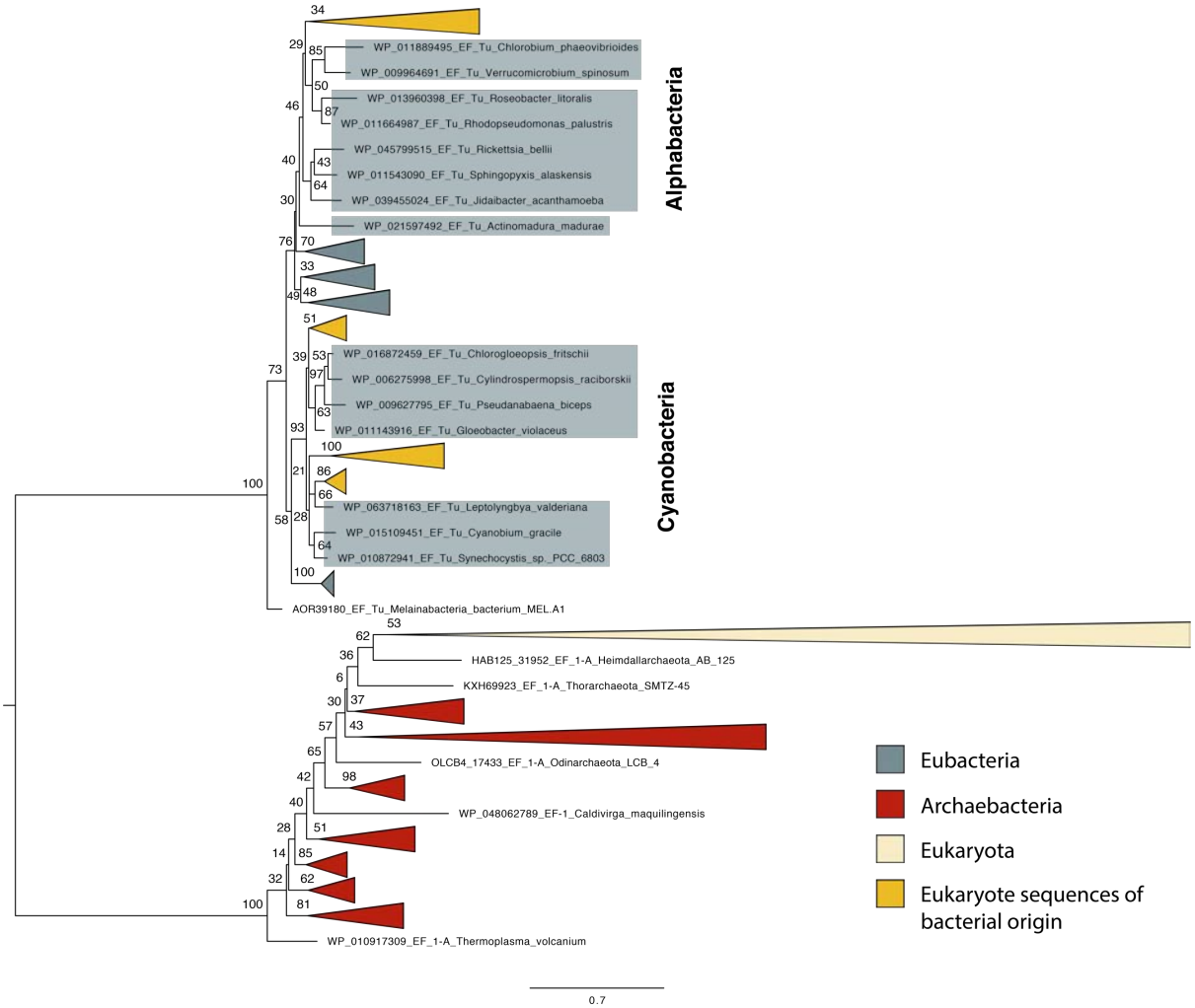


Figure 3.7. Maximum likelihood tree of gene EF-Tu/1. Support values indicated at the corresponding nodes are the Bootstrap Percentage (BP). Branch lengths are the number of character substitutions per site.

3.3.1.4 Histidine biosynthesis subunits A and F

The genes in this family are HisA (1-(5-phosphoribosyl)-5-[(5-phosphoribosylamino) methylideneamino] imidazole-4-carboxamide isomerase) and HisF (imidazole glycerol phosphate synthase, cyclase subunit). Both are involved in histidine biosynthesis where they help catalyse the fourth and fifth parts of the pathway respectively and for a while now it has been thought HisF originated via a duplication event of HisA (Fani et al. 1994). The structure and function of the paralogs is maintained throughout the Bacteria, Archaebacteria and Eukarya. Once aligned, trimmed and cleaned for both erroneous sequences and rogue taxa the gene tree is left with 122 species and 588 sites. Of these the HisF side of the duplication has a topology without any clear domain groupings while the HisA side conforms to an Eocyte tree with eukaryotes diverging within Archaebacteria. The HisA paralog also has a strongly supported root between Eubacteria and Archaebacteria (100 BP) (Fig. 3.8). However, the eukaryotes are sister to a eubacterial species which may be the product of a LGT event. The support for this grouping is not especially high (87 BP). Both subtrees lack a full complement of species from across the domains, instead the eukaryotic diversity is low mostly comprising photosynthetic lineages and fungi with no metazoan taxa present. Due to the issues with the HisF side of the tree we did not use this paralog in the combined dating studies. Additionally, the lack of a LUCA root means that cross-bracing cannot confidently be assigned to date the root of this tree on an individual basis. The HisA paralog was used in the concatenated datasets.

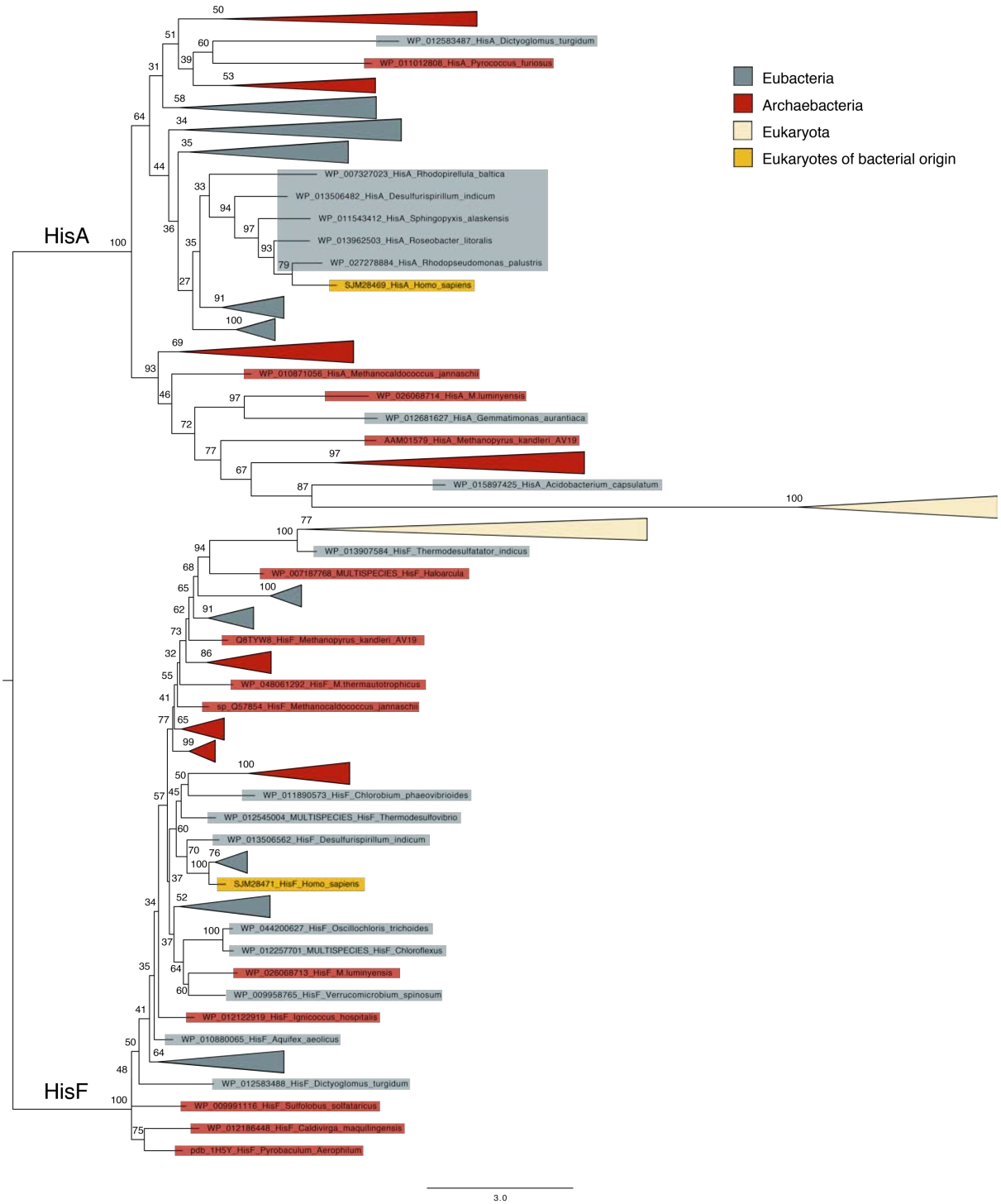


Figure 3.8. Maximum likelihood tree of Histidine biosynthesis subunits A and F. Support values indicated at the corresponding nodes are the Bootstrap Percentage (BP). Branch lengths are the number of character substitutions per site.

3.3.1.5 Ornithine/Aspartate carbamoyltransferases

This family of enzymes is formed of aspartate carbamoyltransferase (ATC) and ornithine carbamoyltransferase (OTC) which catalyse analogous reactions and are respectively involved in the biosynthesis of pyrimidine nucleotides and the conversion of ornithine and carbamoyl phosphate to citrulline (Labedan et al., 1999). The alignment for this gene family contained 177 species. The gene tree presented in Fig. 3.9 exhibits a 3 domains topology in the ATC paralog. This is well supported at all the major nodes, LUCA and the split between Archaebacteria and Eukaryota. There are two eubacterial species branching at the base of Archaebacteria in the ATC paralog. These are likely the result of a LGT. There are no clear domain groupings in the OTC paralog though multiple sequences from Archaebacteria, Eubacteria and Eukaryota are all present (Fig. 3.9). Instead of monophyletic groupings we find a distinct set of eubacterial sequences, along with some photosynthetic eukaryote lineages at all nested within a series of grouped archaebacterial sequences. The root in this paralog is also situated on a branch leading to the Archaebacteria. Once again due to the issues with one paralog, in this case ornithine, we did not use this paralog in the combined dating studies. Additionally, the lack of a LUCA root in this paralog means that cross-bracing cannot confidently be assigned to date the root of this tree on an individual basis. The aspartate paralog was used in the concatenated datasets.

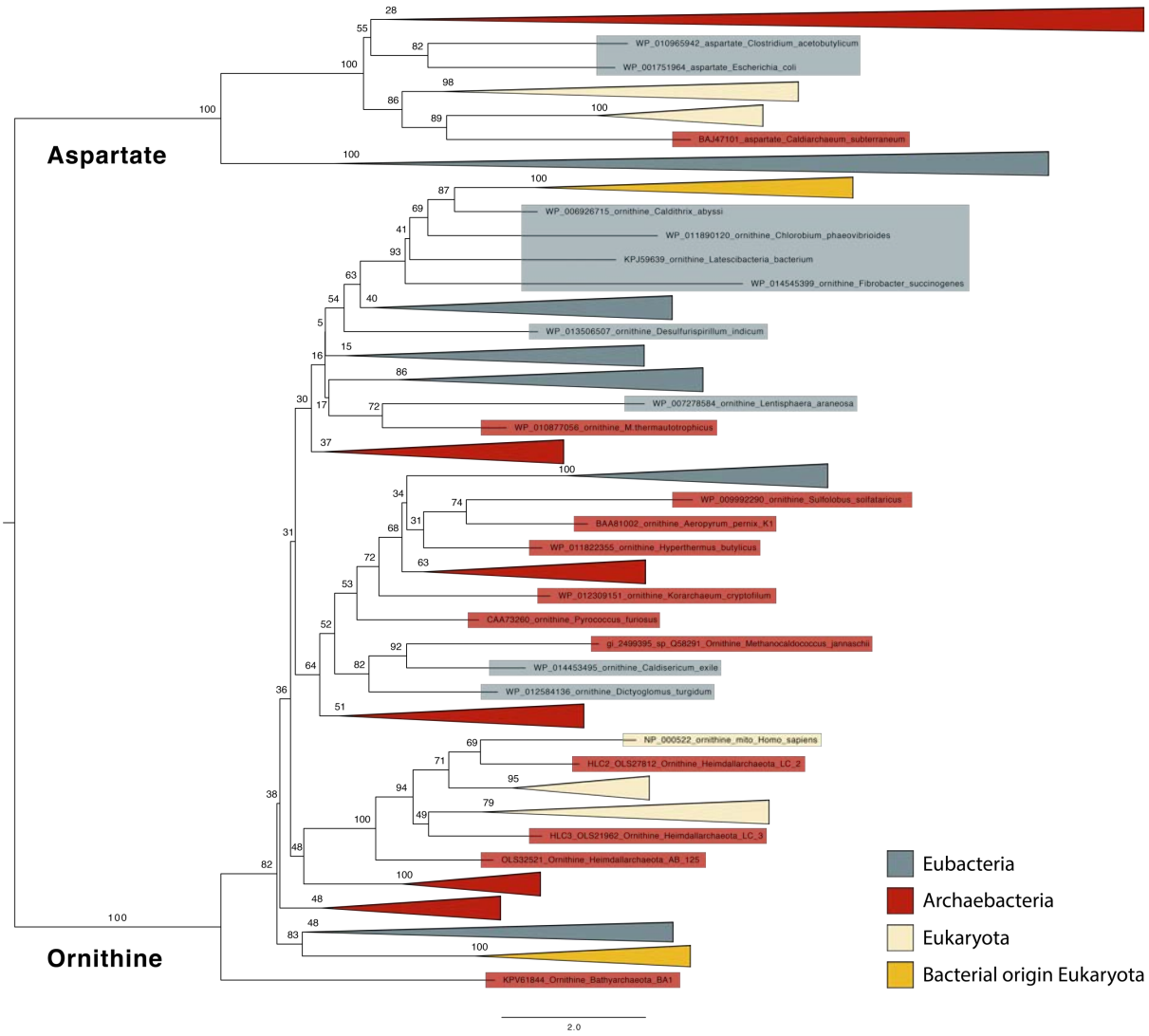


Figure. 3.9. Maximum likelihood tree of Ornithine and Aspartate carbamoyltransferases. Support values indicated at the corresponding nodes are the Bootstrap Percentage (BP). Branch lengths are the number of character substitutions per site.

3.3.1.6 Signal Recognition Proteins

This gene family includes the proteins SRP54(Ffh) and SRα(Ftsy) which are both involved in signal recognition and binding of the ribosome. In our analysis we used an alignment of 143 species and 345 sites which produced a phylogeny with a highly supported split of 100 between Eubacteria and Archaebacteria in both of the paralogs (Fig. 3.10). The placement of eukaryotes conforms to an Eocyte topology in both paralogs as sister lineages to groups of TACK Archaebacteria. However, in both cases this is weakly supported, 36 in SRα(Ftsy) and 29 in SRP54(Ffh). In addition to the eukaryotic sequences of archaebacterial origin, in the SRP54(Ffh) paralog there is a clear grouping of photosynthetic eukaryote lineages with Cyanobacteria. These fall outside the cyanobacterial crown group but within total Cyanobacteria bracketed by *Melainabacteria*. On the whole this gene family has a strong phylogenetic signal with a clear root of the gene prior to LUCA. This makes it a good candidate for dating LUCA via the gene duplication cross-bracing approach. The results of which are detailed in the next paragraph.

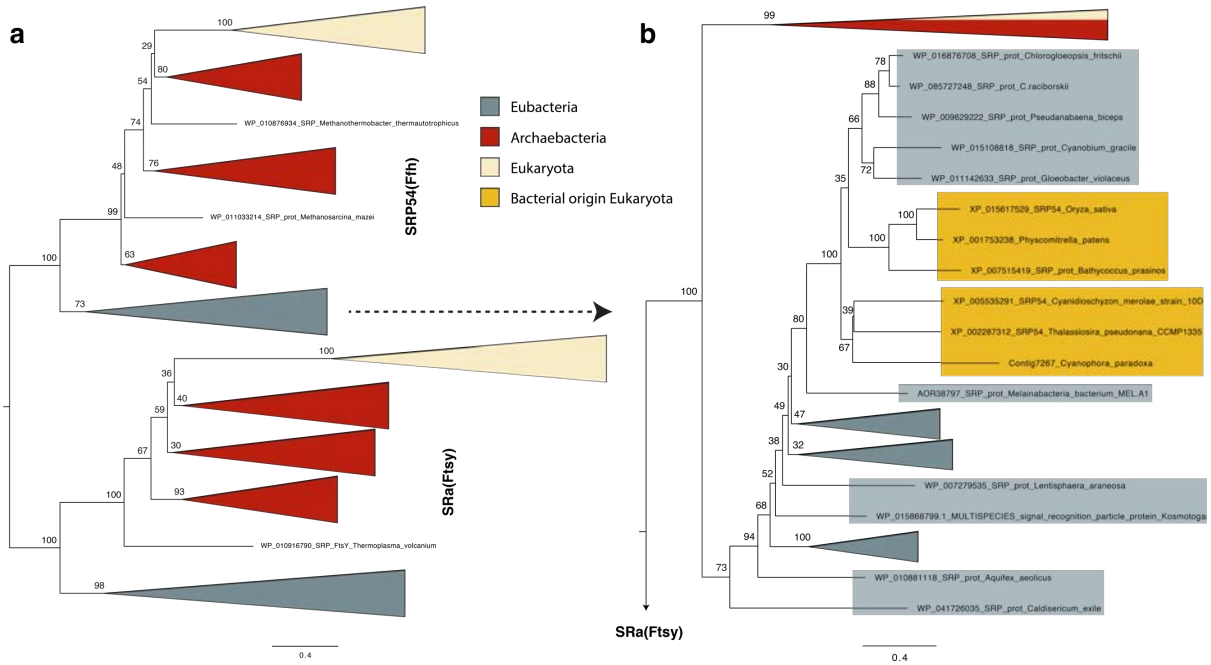


Figure 3.10. Maximum likelihood tree of signal recognition proteins. a) tree containing both SRP54(Ffh) and SRα(Ftsy) and b) a tree focused on SRP54(Ffh) in order to see the chloroplastic sequences. Support values indicated at the corresponding nodes are the Bootstrap Percentage (BP). Branch lengths are the number of character substitutions per site.

3.3.1.7 Tryptophanyl-tRNA and Tyrosyl-tRNA synthetases

This gene family is part of the group of synthetases that attach amino acids to their cognate tRNA's. We used a final alignment of 177 species and 166 sites. When run under maximum likelihood this results in a topology for the tryptophanyl paralog with a basal split between Eubacteria and Archaebacteria with high support (Fig. 3.11a, 100 BP). The eukaryotes resolve within Archaebacteria. This topology is not reflected within the tyrosyl paralog. Here there is a split on one side of most of the archaebacterial species and eukaryotes, and on the other of some more archaebacterial species and all the eubacterial ones. This means the root is placed within Archaebacteria and its placement as such is highly supported (100 BP). However, when a topology for the tyrosyl paralog is inferred in absence of the other half of the gene family and rooted with minimum ancestor deviation rooting (MAD), we find that the root falls between a monophyletic Eubacteria and an Eocyte Archaebacteria plus Eukaryota (Fig 3.11b). This change could be the case for a couple of reasons; firstly that the models we are using cannot accurately capture the true and complex evolution of the gene family, or, that the long branches and considerable evolution between the two gene families is causing the topology to be wrongly resolved in a whole gene phylogenetic production. Within the topology of the tryptophanyl paralog we see clusters of eukaryotes allied with the endosymbiotic lineages. Five photosynthetic species from the green algae and stramenopiles group within crown Cyanobacteria and 3 other eukaryotes branch in two positions within the Alphaproteobacteria. In general, there is high support for this topology within the gene tree but some nodes, especially towards the tips, display much lower support of down to ~50.

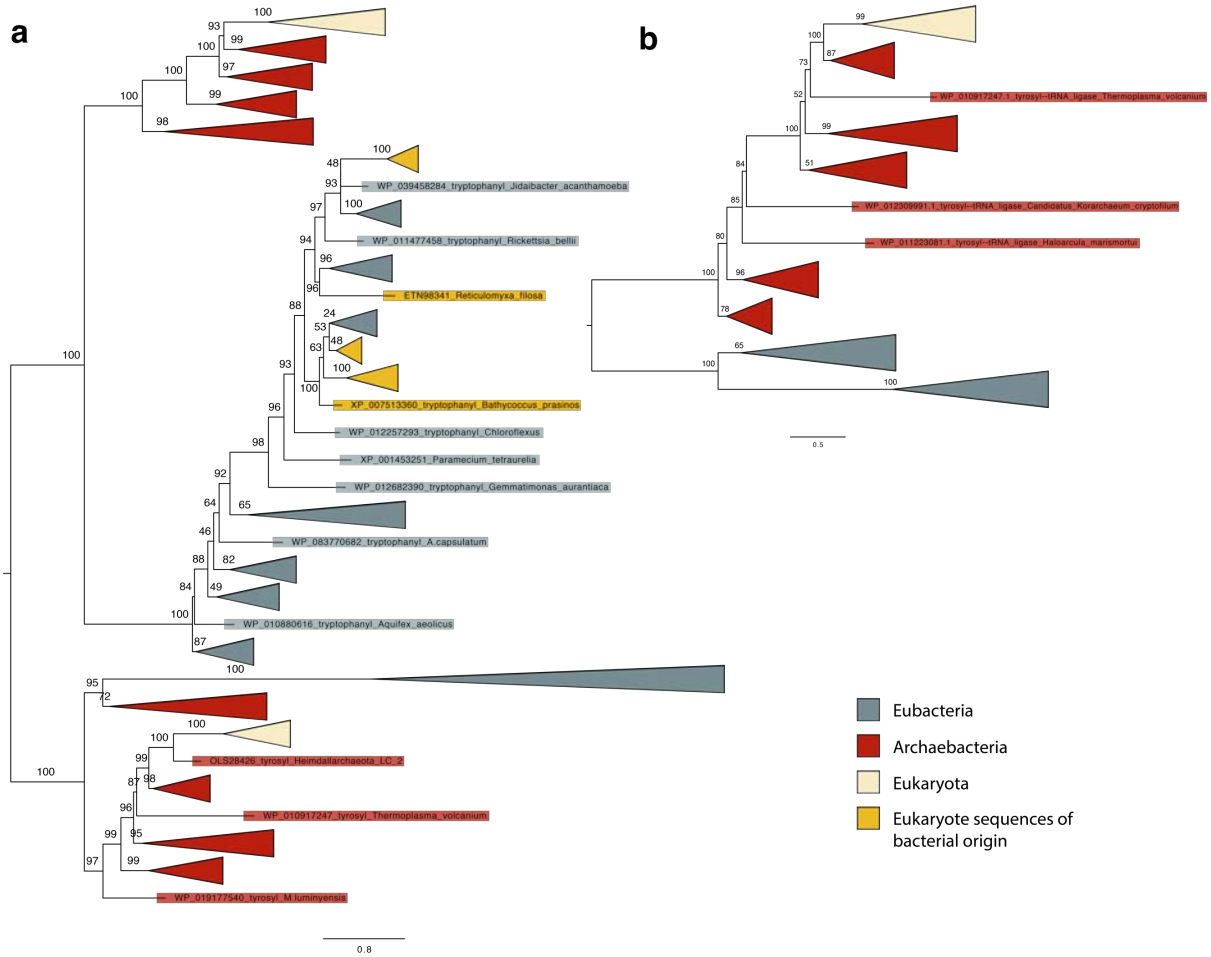


Figure 3.11. Maximum likelihood tree of Tryptophanyl-tRNA and Tyrosyl-tRNA synthetases. a) A complete tree with both genes Tryptophanyl-tRNA above and Tyrosyl-tRNA below and b) a maximum likelihood tree of Tyrosyl-tRNA synthetase which has been rooted using MAD. Support values indicated at the corresponding nodes are the Bootstrap Percentage (BP). Branch lengths are the number of character substitutions per site.

3.3.1.8 Valyl-, Methionyl-, Isoleucyl- and Leucyl- tRNA synthetase

This gene family has previously been used to look at the topology of the tree of life, especially with regard to the placement of eukaryotes (Brown and Doolittle 1995). Previous analyses have looked at three paralogous genes valyl- (ValRS), isoleucyl- (IleRS) and leucyl- (LeuRS) tRNA synthetase which are part of the group 1 aminoacyl-tRNA synthetase genes. The BLAST search performed as part of this work identified methionyl- (MetRS) as an additional gene in this gene family. MetRS was found to be the outgroup with the other three genes grouping as ValRS and LeuRS most closely related with IleRS (Fig. 3.12). The support for these nodes ranges from non-significant to high (between 83 and 100). Within the different paralagous genes we find topological inconsistencies when compared with recognised species phylogenies, especially in the IleRS paralog where there is no clear domain monophyly for the eubacterial and archaebacterial species included. The other three genes preserve monophyletic groupings of the two main lineages with the exception of 2 eubacterial sequences within the archaebacterial part of the MetRS gene family. Both MetRS and ValRS have some eukaryotic sequences of eubacterial origin, in MetRS these clearly group within Cyanobacteria. However, neither have any eukaryotic sequences within Archaebacteria. Eukaryotic sequences arising from within Archaebacteria are observed in the LeuRS gene.

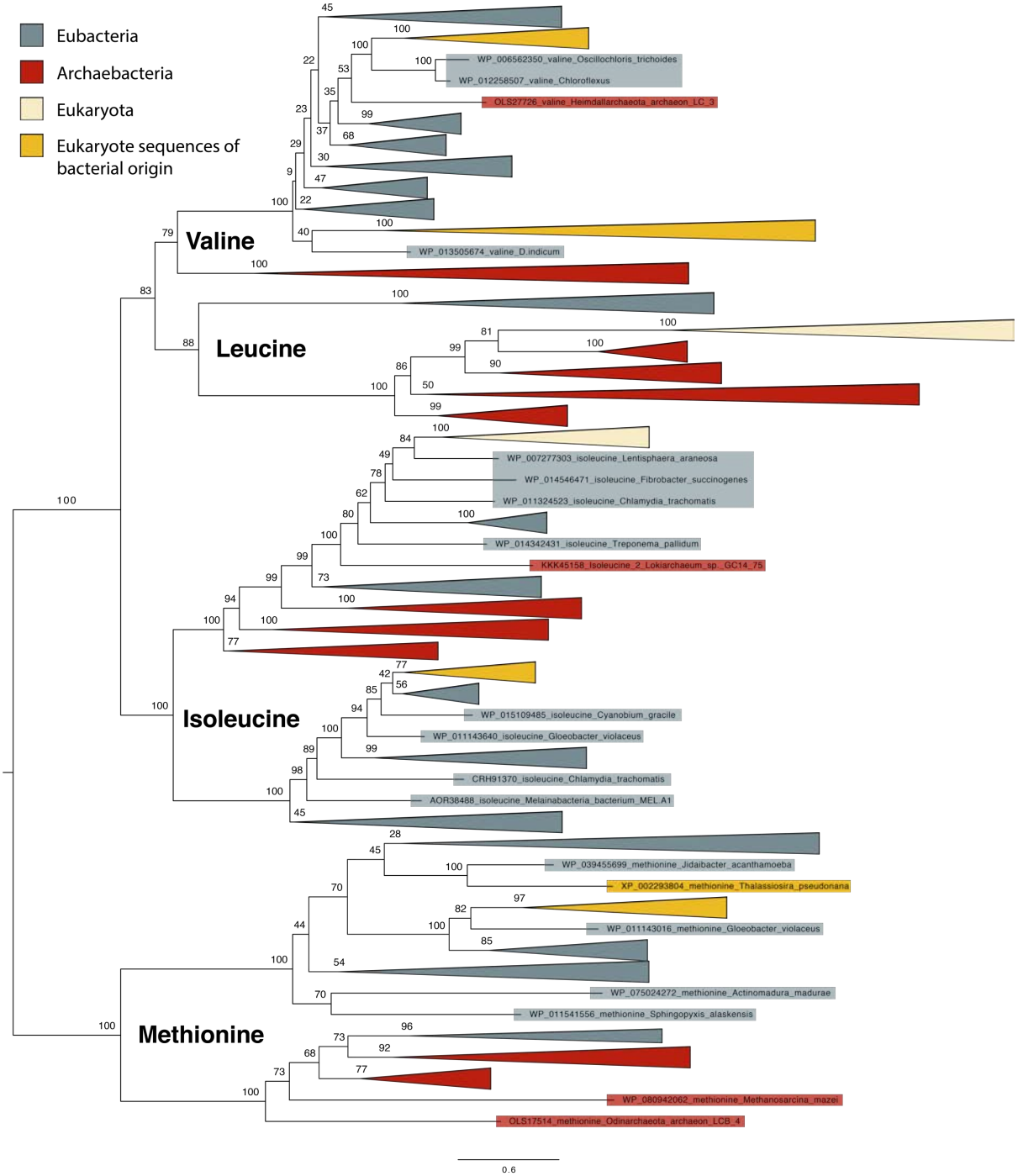


Figure 3.12. Maximum likelihood tree of the genes Valyl-, Methionyl-, Isoleucyl- and Leucyl- tRNA synthetase. Support values indicated at the corresponding nodes are the Bootstrap Percentage (BP). Branch lengths are the number of character substitutions per site.

3.3.1.9 Concatenated analysis

Of the 8 genes we examined 7 had enough information to be concatenated and analysed as part of a larger dataset. This analysis was purely to examine the topology of the tree, rather than looking at the duplication event. This meant that we could make use of the combined information provided by the individual genes. The gene families were separated into individual paralog files before being concatenated such that each paralog was part of the concatenation in its own right (see Fig. 3.13). As the individual paralogs all have the same root, LUCA, they should in theory have originated at the same time. Meaning that the speciation event that led to the ancestors of Eubacteria and Archaebacteria should have occurred at the same time in each gene. The resulting tree produces a topology where Archaebacteria and Eukaryota are more closely related to each other than either are to Eubacteria (Fig. 3.14). Specifically, the eukaryotes branch next to the Asgardarchaeota though without high support (0.82 PP). Within Alphaproteobacteria the mitochondrial sequences group next to the Rhodopseudomales. However, *Rickettsia* and *Jidaibacter* are not found to group with the other alphaproteobacterial lineages (perhaps due to LGT in some of the concatenated genes). Therefore, it is hard to assess whether the grouping of Rhodopseudomales with the mitochondria is informative about the affinities of the mitochondria within Alphaproteobacteria. In Cyanobacteria the chloroplast sequences divergence at the base of the group in a highly supported position (1 PP). *Melainabacteria* is the outgroup to both the cyanobacterial and chloroplastic sequences.

3.3.1.10 Duplicate concatenated analysis

For this analysis the gene families were concatenated such that the gene family alignments were kept intact. Meaning that unlike the previous analysis the duplicate sets of paralogs within each family were kept together (see Fig. 3.13). Hence, the duplication event at the base of the tree is retained and we can examine the topology produced by the combined information from all of the genes. The topology produced is similar on both sides of the duplication (Fig. 3.15) with the root falling between Archaebacteria and Eubacteria with a support of 100 BP. Eukaryotes emerge from within Archaebacteria as sister to all Heimdallarchaeota species in duplicate 1 and

Heimdallarchaeota_archaeon_AB_225_2 and Heimdallarchaeota_archaeon_LC_2_2 in duplicate 2. Though in both cases the support for these specific groupings is low (55 and 78 BP respectively). In duplicate 1 the mitochondrial sequences group at the base of Alphaproteobacteria and the same is true in duplicate 2. For the chloroplastic sequences, in duplicate 1 they branch within crown Cyanobacteria and in duplicate 2 they branch basally to the other cyanobacterial species.

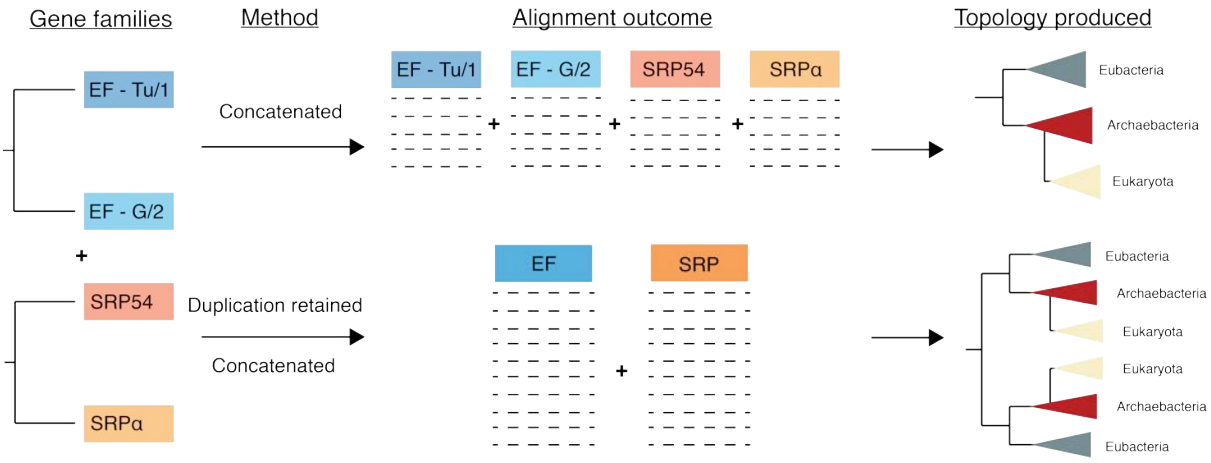


Figure 3.13. Illustration of the two concatenation methods used to combine information from the gene families. The first method involved separating the two paralogs from within each gene family so that they could be concatenated in their own right. This produced a topology with one LUCA node which is at the root of the tree. The second method involved concatenating the gene families with both paralogs in the same alignment file. This resulted in a topology with two LUCA nodes and a pre-LUCA duplication node. In this figure the elongation factor (EF) and signal recognition particle (SRP) gene families have been used as examples.

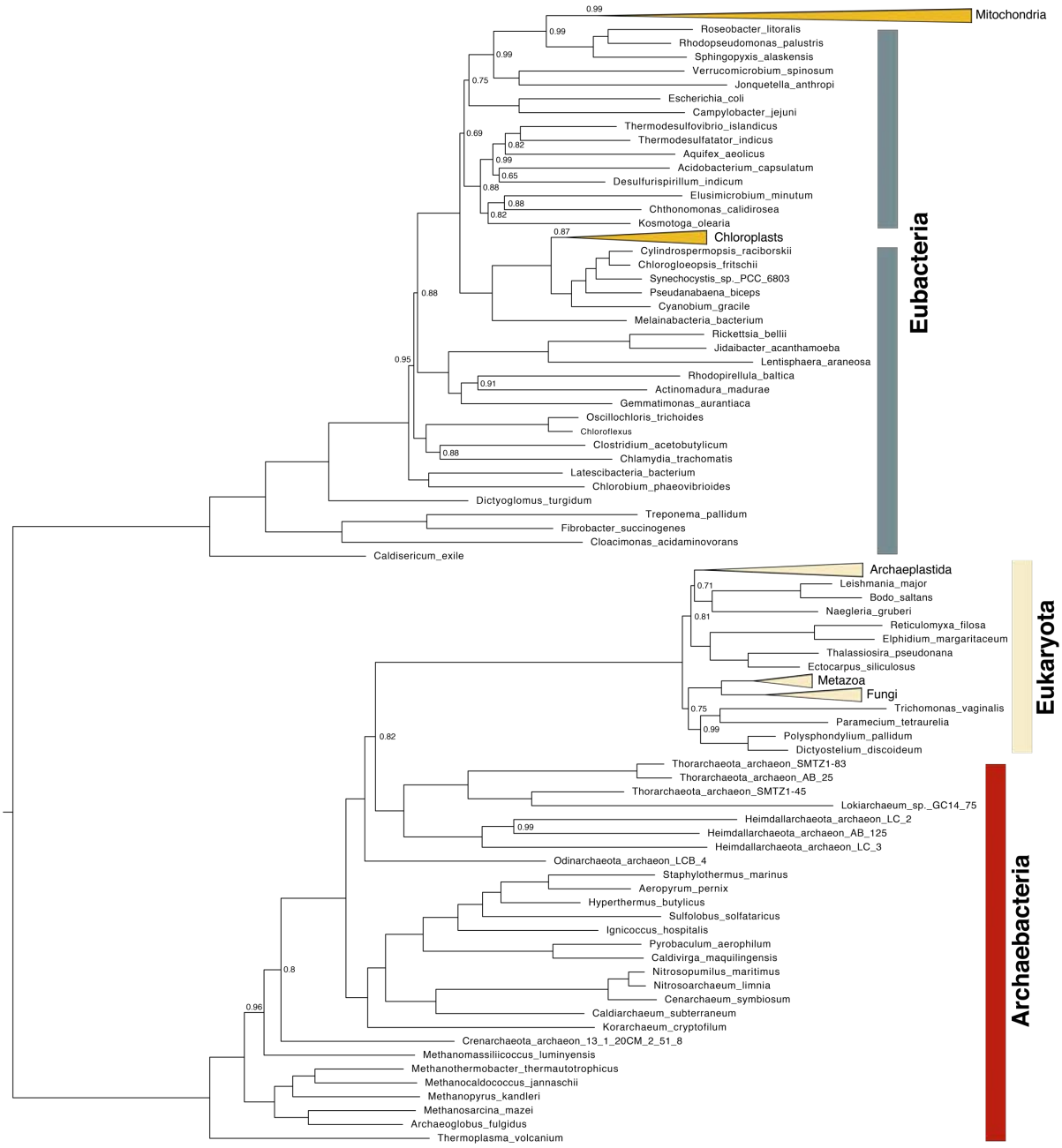


Figure 3.14. PhyloBayes tree of concatenated genes. Node supports are the posterior probability of the node when the support is less than 1.

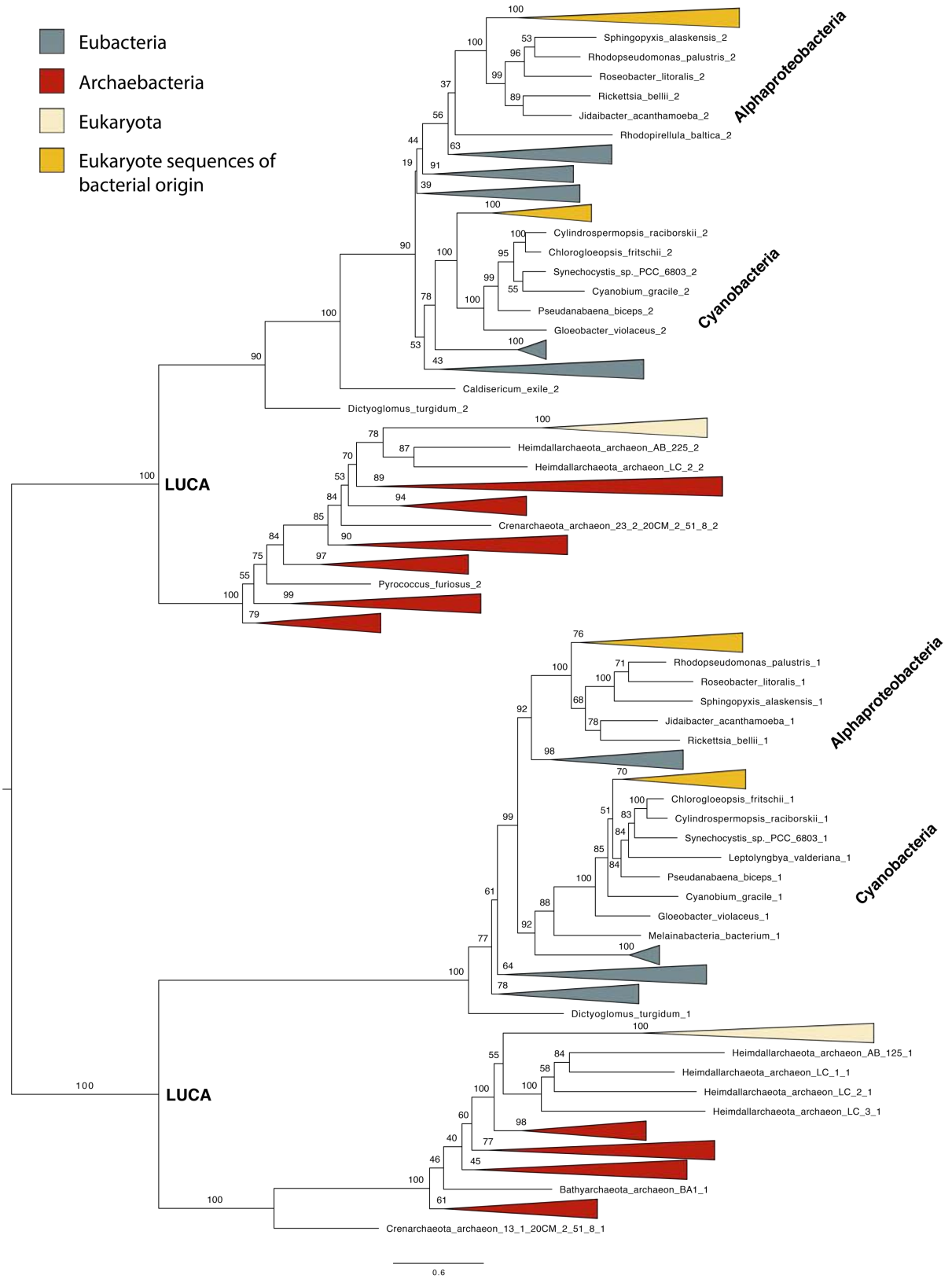


Figure 3.15. Maximum likelihood phylogeny of concatenated genes with the duplication retained. Support values indicated at the corresponding nodes are the Bootstrap Percentage (BP). Branch lengths are the number of character substitutions per site.

3.3.2 Divergence dates

As with the topology analyses the gene trees share similar divergence dates and their divergence time trees have a number of common properties. When used to estimated divergence times all 5 gene trees which possessed a LUCA root produced similar ages for this node (Fig. 3.16). The ATPase gene family finds an age for LUCA close to the age of the Earth between ~4.4 Ga and 4.52 Ga. This age is recovered both when LUCA is the root node and when there are two LUCA nodes and the analysis is cross-braced (See Appendix Fig. A.12). In both cases the credible interval for this node is small. The age of LUCA in the EF gene tree is around 4.26 – 4.44 Ga and in the signal recognition protein (SRP) divergence time tree LUCA is dated to between 4.23 – 4.46 Ga. We dated a tree with a tyrosyl root falling between Eubacteria and Archaebacteria. In this case LUCA diverges ~4.36 - 4.517 Ga consistent with the other genes. Finally, the Val-Leu-MetRS age for LUCA is slightly younger than the other genes ~ 4.16 – 4.3 Ga. The ages produced by the single gene trees overlap and amount to a total interval for LUCA of 4.16 – 4.517 Ga. The consistency between these ages can be seen in Fig. 3.16 where all the genes date LUCA to prior to 4.0 Ga.

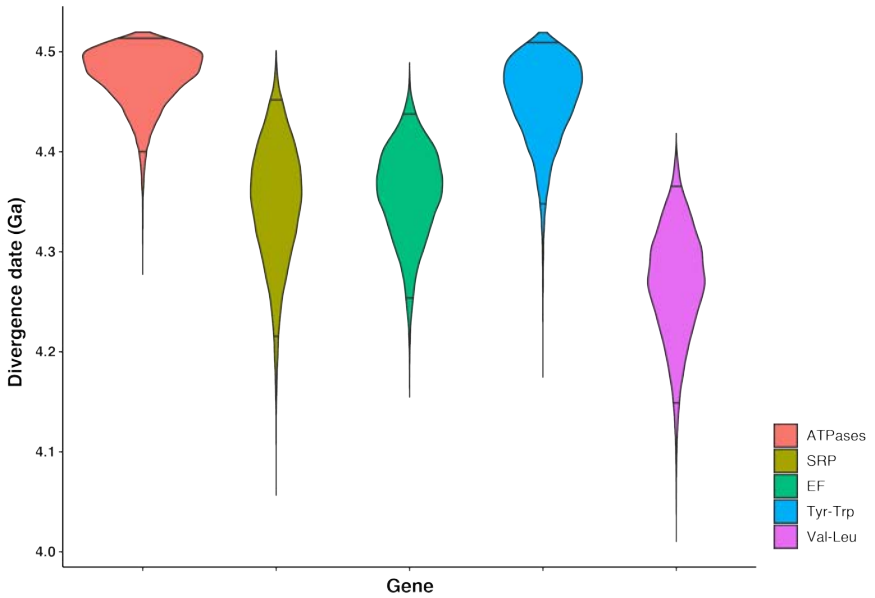


Figure 3.16. Violin plot of posterior age estimates for the last universal common ancestor from 5 of the gene trees. The black lines in each plot indicate the 95% credible interval and the width is an indicator of how the values are distributed. In the key; ATPases is F-type and V-type ATPases, SRP is signal recognition proteins, EF is elongation factors EF-G/2 and EF-Tu/1, Tyr-Trp is Tyrosyl-tRNA and Tryptophanyl-tRNA synthetase and finally Val-Leu is Valyl-, Leucyl- and Methionyl- tRNA.

Similar to the LUCA node other divergences of interested, such as crown Eubacteria and crown Archaebacteria, also exhibit a large amount of overlap between the divergence dates produced by the individually dated gene trees (Fig. 3.17). The age of both crown groups is slightly younger in the Val-Leu-MetRS gene family than in the other genes. The youngest date for Eubacteria is 3.49 Ga in LeuRS and the greatest is ~4.516 Ga in Tyrosyl-tRNA. EF-Tu/1 produces the youngest date for Archaebacteria ~3.64 Ga and Tyrosyl-tRNA produces the oldest ~4.519 Ga. In both cases this means the overall 95% credible intervals stretch from just prior to 3.5 billion years ago up to 4.5 billion years ago. Although these nodes are not cross braced the ages between the two sides of the duplication, indicated in Fig. 3.17 by the labels _1 and _2 show good agreement with each other. These ages are close to the age of the Earth.

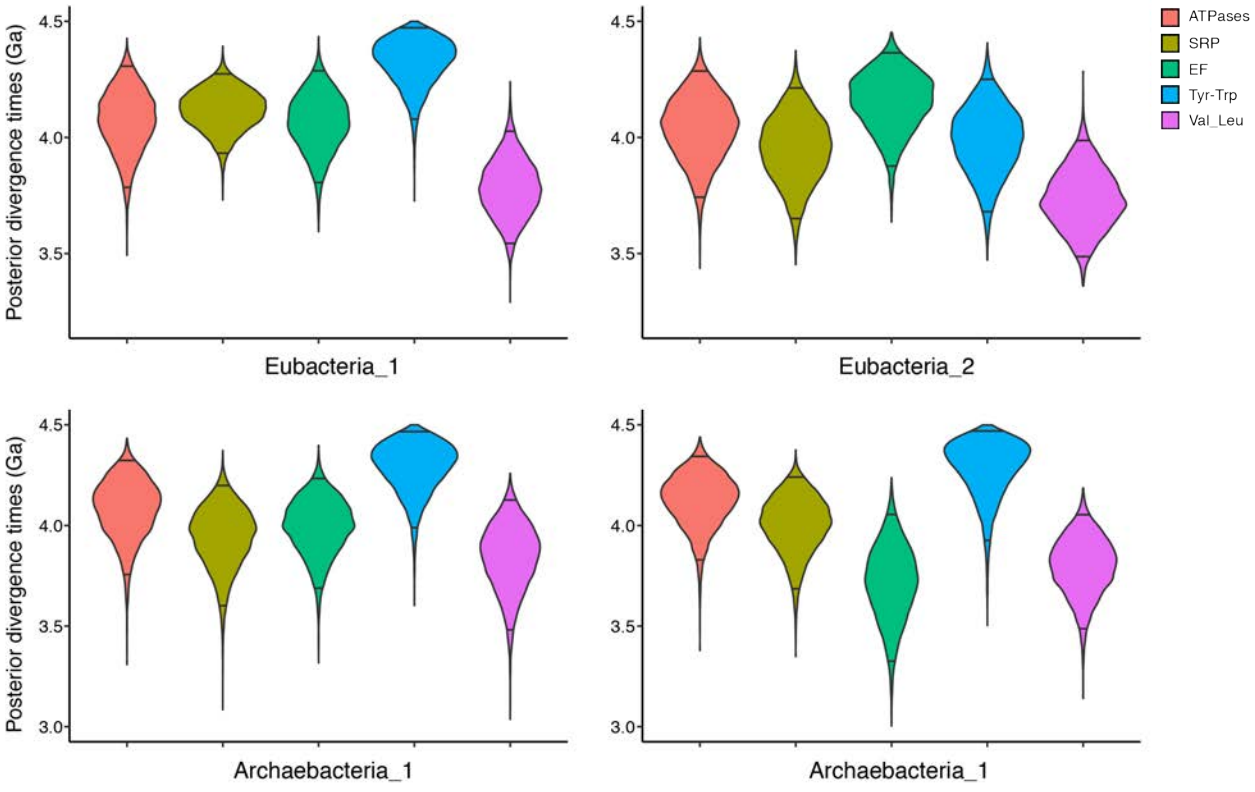


Figure 3.17. Violin plots of posterior age estimates for the nodes crown Eubacteria and crown Archaebacteria from individually dated gene trees. The black lines in each plot indicate the 95% credible interval and the width is an indicator of how the values are distributed. In the key; ATPases is F-type and V-type ATPases, SRP is signal recognition proteins, EF is elongation factors EF-G/2 and EF-Tu/1, Tyr-Trp is Tyrosyl-tRNA and Tryptophanyl-tRNA synthetase and finally Val-Leu is Valyl-, Leucyl- and Methionyl- tRNA.

Not all of the paralogs have a distinct group of eukaryotes nested within Archaebacteria. However, the ages for the paralogs that do can be seen in Figure 3.18 and there is at least one in each of the dated gene trees. The credible intervals where this node is present are wide and vary between the genes. The oldest found the ATPases and the youngest in the elongation factor genes. In general, the divergence dates produced for eukaryotes are between 2 and 3 billion years ago with credible intervals stretching over 1 billion years.

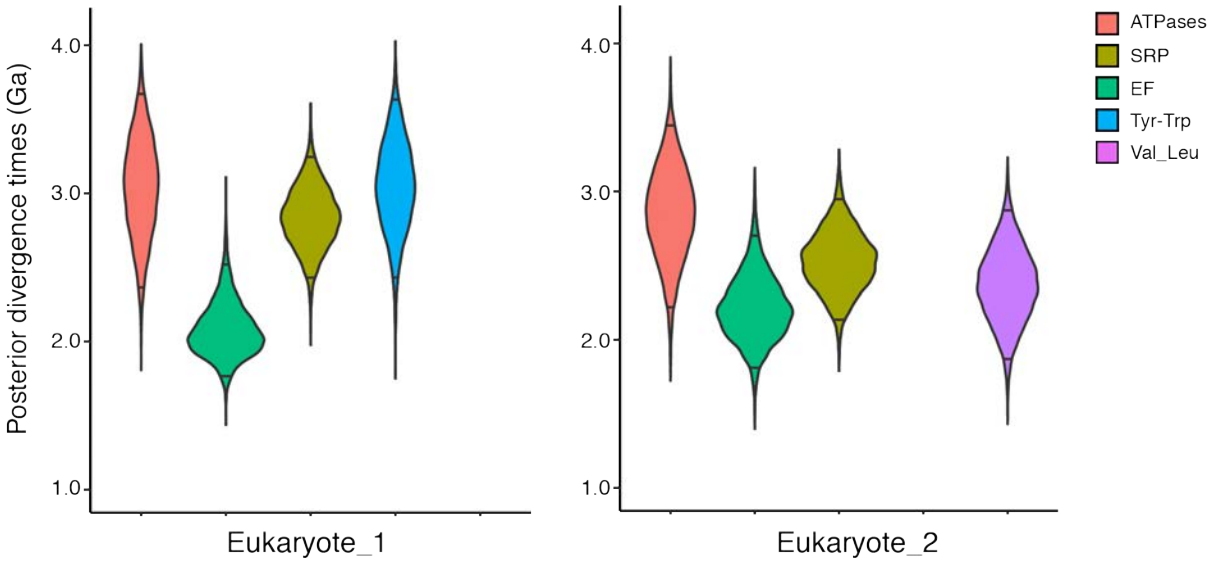


Figure 3.18. Violin plots of posterior age estimates for crown Eukaryota from individually dated gene trees. The black lines in each plot indicate the 95% credible interval and the width is an indicator of how the values are distributed. In the key; ATPases is F-type and V-type ATPases, SRP is signal recognition proteins, EF is elongation factors EF-G/2 and EF-Tu/1, Tyr-Trp is Tyrosyl-tRNA and Tryptophanyl-tRNA synthetase and finally Val-Leu is Valyl-, Leucyl- and Methionyl- tRNA.

An example of a single gene tree can be viewed in Figure 3.19. The wide credible intervals displayed using blue bars can be clearly seen. The intervals are often smaller when a node possesses a calibration, especially towards the tips of the tree such as the split between *Physcomitrella patens* and the angiosperm taxon included, *Oryza sativa*. The timescale along the bottom is measured in millions of years before present and clearly shows that LUCA is very close the age of the Earth and the moon forming impact, the maximum age upon this timescale. The credible intervals for crown Eubacteria, Archaebacteria, Alphaproteobacteria and Cyanobacteria are also wide but consistent across the duplication. The format of this gene tree is representative of all the individually dated genes.

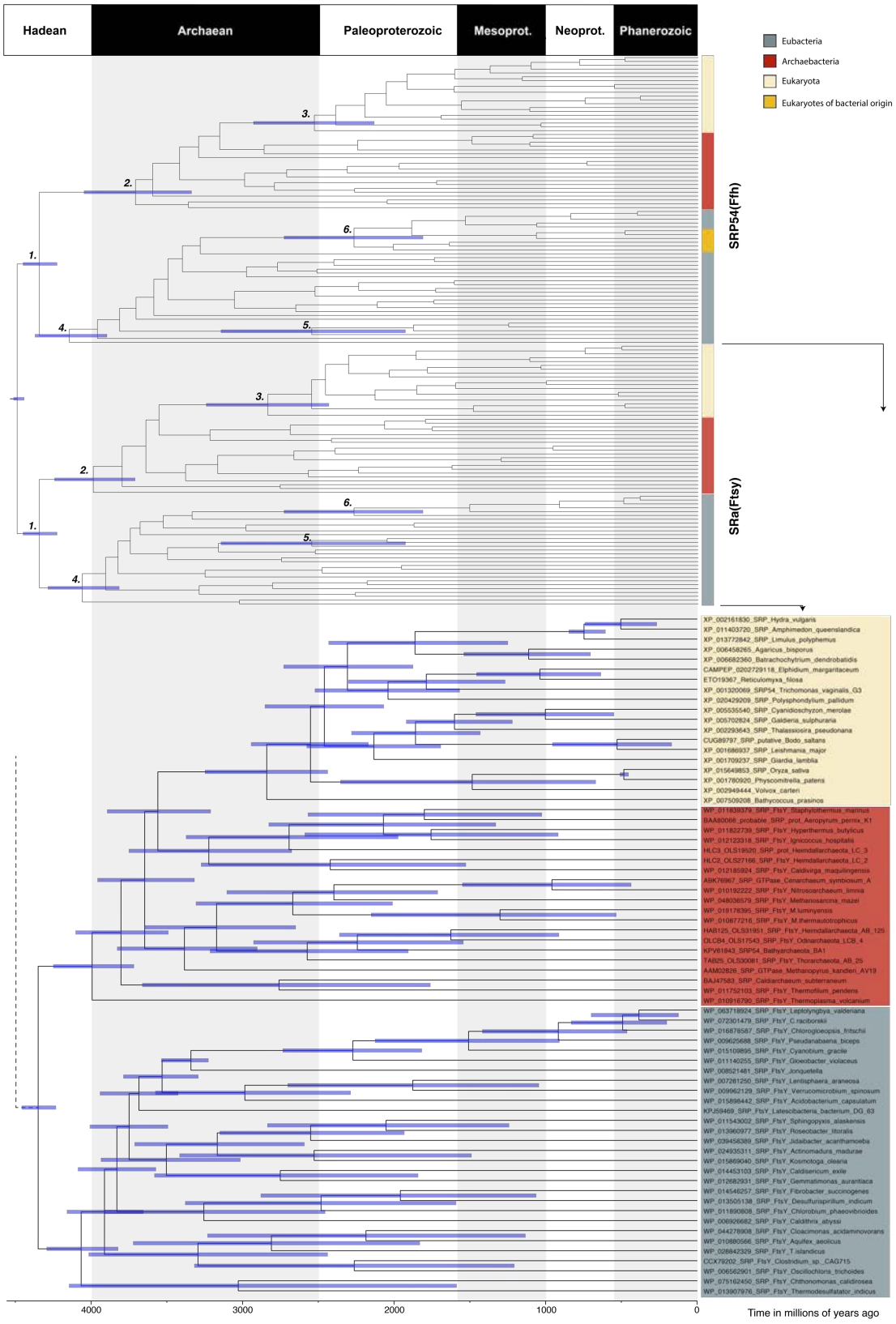


Figure 3.19. Divergence time tree for the signal recognition protein gene tree. In the lower half of the tree the SRP(Ftsy) gene has been highlighted for a better view of the confidence intervals. The blue bars represent the 95% credible intervals. In the top tree some key nodes have been highlighted. These are; 1 LUCA, 2 LACA (last archaeal common ancestor), 3 LECA (last eukaryotic common ancestor), 4 LBCA last bacterial common ancestor, 5 crown Alphaproteobacteria and 6 crown Cyanobacteria.

Ultimately the duplicate retained concatenated analysis could not be cross-braced. Though this was attempted, the software produced unlikely results pushing the origin of life prior to the formation of the solar system (this tree can be viewed in the appendix, Fig. A.13), this is likely because of a bug in the version of MCMCtree we used which has been explicitly developed by Prof. Z. Yang (a collaborator on this project) to be tested as part of this study. Instead of using this we therefore cross-calibrated the tree. These results are presented in Fig. 3.20. On one side of the duplication LUCA dates to between 3.89 and 4.18 Ga. On the other LUCA dates to between 4.39 and 4.51 Ga. This difference means that the ages of these two nodes do not overlap. Hence, we suggest that to two ages should be combined producing a conservative estimate for the age of LUCA between 3.89 to 4.51 Ga. This is similar to the ages produced by the individual gene trees (4.16 – 4.517 Ga, Fig. 3.16).

Ages for crowns nodes display variable amounts of overlap, less for crown Archaebacteria and crown Eukaryota, more for crown Eubacteria, Alphaproteobacteria and Cyanobacteria. If we employ the same conservative approach as above, we find dates for these nodes as; Eubacteria 3.57 – 3.89 Ga, Archaebacteria 3.25 – 4.49 Ga, Eukaryota 1.79 – 2.53 Ga, Alphaproteobacteria 1.47 – 2.33 Ga and Cyanobacteria 1.2 – 2.23 Ga. These nodes are named 4, 2, 3, 5 and 6 in Fig. 3.20. The credible intervals are still relatively large when compared to the individually dated gene trees. It will be interesting to see how the cross-braced results compare with the averages presented above as soon as the problems with the current version of MCMCtree are resolved.

Divergence time trees for all individual genes can be found in Appendix A (Fig. A.1 – A.12)

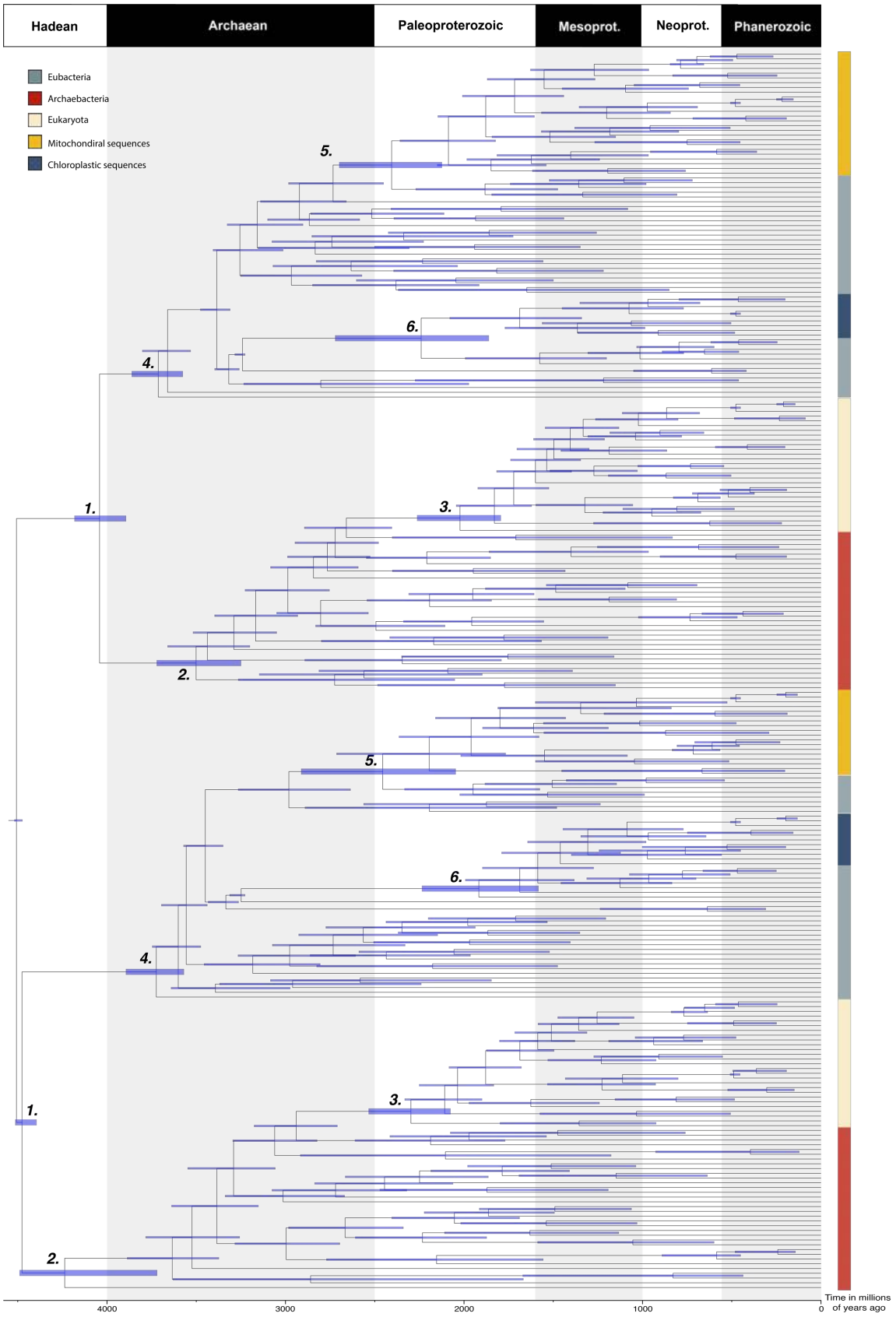


Figure 3.20. Divergence time tree for concatenated genes with duplication retained. The blue bars represent the 95% credible intervals. Some nodes of interest have been highlighted. These are; 1 LUCA, 2 LACA (last archaeal common ancestor), 3 LECA (last eukaryotic common ancestor), 4 LBCA last bacterial common ancestor, 5 crown Alphaproteobacteria and 6 crown Cyanobacteria.

3.4 Discussion

3.4.1 Topology

Results from across the 8 gene families reveal some common properties, in all cases the topologies of the gene trees are different to an expected species tree, something which is unsurprising given the nature of how genes evolve (Degnan and Rosenberg 2009). In general, the gene trees preserve some evidence of domain groupings, though rarely with a monophyletic Archaebacteria or Eubacteria. Additionally, as mentioned in the results, 3 of the gene families have at least one paralog which does not exhibit any strong domain monophyly or a distinct LUCA root, the HisAF family, the CPS family and the OTC/ATC family. An Eocyte topology is found in all cases, as it has been by numerous other studies working with different kinds of data (Lake et al. 1984; Rivera and Lake 1992; Pisani, Cotton, and McInerney 2007; Cox et al. 2008; Williams et al. 2013; Williams et al. 2012; Spang et al. 2015; Zaremba-Niedzwiedzka et al. 2017). However, when an Ecoyte topology is produced eukaryotes are not always the sister lineage to Asgardarchaeota (Spang et al. 2018; Spang et al. 2015; Zaremba-Niedzwiedzka et al. 2017), but to a variety of other archaeal lineages such as in the SRP and Tyr-Trp gene trees. Despite the variation within the individual gene trees in both the concatenated analyses, where the genes are individually concatenated and where the duplicated node is preserved, the eukaryotes do resolve as sister to Asgardarchaeota.

The topology within the domains lacks consistency between the gene trees. Each resolves a slightly different placement for the component lineages. This is true of the position of mitochondrial and chloroplastic sequences which fall in a variety of places both within and at the base of their respective closest relatives, Alphaproteobacteria and Cyanobacteria. In general, the gene trees suggest that the endosymbiotic transfer events occurred from a now extinct lineage along the stem of the groups. This is a more basal position than earlier analyses have produced, with the chloroplastic eukaryotic sequences often appearing above the basal cyanobacterial lineages, *Gloeobacter* in addition to some *Synechococcus* species, but below below the bulk of cyanobacterial diversity (Criscuolo and Gribaldo 2011; Sánchez-Baracaldo et al. 2017; Shih and Matzke 2013). The ancestral mitochondrion is argued

to have arisen from a number of different alphaproteobacterial lineages (Wang and Wu 2014, 2015; Abhishek et al. 2011; Fitzpatrick, Creevey, and McInerney 2005; Thiergart et al. 2012) and recently there have been suggestions that it was most closely related to a group somewhere along the alphaproteobacterial stem (Martijn et al. 2018; Esser et al. 2004; Rodríguez-Ezpeleta and Embley 2012). Our gene trees seem to confirm this idea. Further support for this is provided by the concatenated datasets where the mitochondrial and chloroplastic lineages also fall at the base of Alphaproteobacteria and Cyanobacteria respectively. The support for many nodes is high in both the individual gene trees and the concatenated analyses (Fig. 3.3 – 3.15). Crucially this is true for those nodes of key interest such as the domains, endosymbiotic lineages, and those with calibrations. However, in each gene tree there are also nodes with very low support.

The varying position of the root within the F-type and V-type ATPases is a known phenomenon likely due to problems with analysing very anciently diverging genes (Gogarten et al. 1989). The final topological analysis presented here for this gene family finds a grouping of the two F-type subunits and a grouping of the two V-type subunits (Fig. 3.3). However, in general the root is most often found between the groupings of F-type subunit B plus V-type subunit A and F-type subunit A plus V-type subunit B (Shih and Matzke 2013; Zhaxybayeva, Lapierre, and Gogarten 2005; Mulkidjanian et al. 2007). Thus, producing dates for a gene tree with this root seemed the most logical step. Likewise, previous studies have found that the root in the SRP gene family lies firmly between the Bacteria and a more closely related Archaea and Eukarya (Gribaldo and Cammarano 1998) in the SRα(Ftsy) gene. Though support for an Eocyte tree vs. a three domains tree is not consistent (Philippe and Forterre 1999; Gribaldo and Cammarano 1998). In the Tryptophanyl-tRNA and Tyrosyl-tRNA synthetase gene family Tyrosyl-tRNA has previously been demonstrated to follow a traditional 3 domains tree of life and Tryptophanyl an Eocyte topology with eukaryotes grouping within Archaebacteria (Kollman and Doolittle 2000). The analyses presented here show that some genes in the tyrosyl side of the tree might be subject to some level of biases or long branch attraction causing the root in this tree to vary (Fig. 3.11a). Another, more recent maximum likelihood phylogeny for the paralog has also recovered an Eocyte topology (Furukawa et al. 2017).

Labdean and colleagues (1999) found support for a 2 domains topology within the ATC family, as produced here, but within the OTC family they report two main groups of OTC sequences, one containing only bacterial sequences and the other containing a mixture of bacterial, archaeal and eukaryotic sequences. A more recent analysis proposed that there are 3 groupings of OTC sequences with the 3rd group being formed of fungal and metazoan species (Zúñiga, Pérez, and González-Candelas 2002). The tree reported here (section 3.3.1.5, Fig. 3.9) does not reflect either of these topologies within the OTC paralog but it does share the characteristic that domain monophyly is not preserved and that there is no clear LUCA root. Thus, this gene is not useful for investigating the age of LUCA.

The variation between the gene trees could have been produced by a few different factors. Firstly, the presence of LGT resulting in the placement of eubacterial and archaebacterial sequences within another domain. This is included in the overarching issue of genuinely different gene histories which could also include losses specific to certain gene families. Each of the genes will have taken a slightly different path over the course of their >4 billion years of evolution. The differences could also be due to stochastic error and problems with modelling deep divergences. Although we strove to use the best models for each gene, and the most appropriate available, they may still fail to capture the full complexity of the evolution of each gene. Where LGTs were present they were retained in topological analyses and divergence time estimations for individual gene trees in order to see when the possible transfer events may have occurred. However, they were removed when looking at concatenated topology and divergence dates to minimise error within these datasets as the LGT events do not conform to the species tree vertical transfer. The analyses undertaken here included a larger number of archaeal sequences and often more eukaryotic sequences than used in previous studies. Perhaps of most importance this includes the recently discovered and described Asgardarchaeota (Spang et al. 2015; Zaremba-Niedzwiedzka et al. 2017). This allowed a more in depth look at where the root of these genes lies and whether or not they can actually be used, firstly to investigate a root for the tree of life, and secondly whether they can be used to date a tree of life. The concatenated topologies confirm that the combined gene trees have a strong signal which more or less lines up with what we would expect for the species tree. In some cases, there is variation at the tips but the expected overall pattern of a LUCA root between

the monophyletic Archaebacteria and Eubacteria is still produced. Eukaryotes emerge as sister to Asgardarchaeota (Spang et al. 2018; Spang et al. 2015; Zaremba-Niedzwiedzka et al. 2017) in some genes. This is confirmed by analysis of the concatenated genes.

3.4.2 Divergence dates

Despite the topological variation between gene trees in general we find divergence dates in common with previous studies for some of the major nodes (Betts et al. 2018; Eme et al. 2014; Parfrey et al. 2011; Sánchez-Baracaldo et al. 2017; Schirrmeister et al. 2013). The spread of ages for LUCA from genes with two paralogs containing a LUCA root can be seen in Fig. 3.16. All of these results overlap, though to varying degrees, and date LUCA to older than 4.16 Ga with the upper age estimate extending close to the age of the Earth. In all cases this places LUCA as older than the late heavy bombardment (~3.8-4.1 Ga). Cross calibration using the concatenated dataset generates a tree (Fig. 3.20) where there is variation between the ages of nodes that possess calibrations. This means that the LUCA node has age ranges which do not overlap. Despite this, both of the nodes are still very ancient and produce a combined age of 3.89 to 4.51 Ga. Cross-calibration provides more information than if LUCA was being dated as the root node of the analysis. This coupled with the information from the individual gene trees corroborates the age for this ancient node and underlines why the molecular clock approach is so important for this early time period. The rock record by these ages is non-existent, other than a small number of zircons (Harrison, Bell, and Boehnke 2017), and this methodology can help us in elucidating the timescale of life for very early lineages. It backs up suggestions that LUCA was a very ancient organism on our planet and that life survived huge amounts of environmental upheaval early in its evolution. In the gene trees, nodes which either have a calibration near the tips, or a calibration which is quite strict, have small 95% highest posterior density (HPD) credible intervals (For example Fig. 3.19). However, for other nodes the HPD widths are large, often covering in excess of one billion years, for example the node labelled 5 (Fig. 3.19), crown Alphaproteobacteria, has a very wide credible interval. This is likely to be an effect of gene-level analyses where, although they can be used in dating explorations, the information accessible from one gene is not enough to produce more precise time estimates. The lack of precision in these estimates means that we are likely capturing the accurate date

for these divergences but that it is harder to infer anything about the evolutionary timescales of these lineages.

The large credible interval widths are found on nodes including some of interest such as the origin of crown Eubacteria and crown Archaebacteria. These nodes have respective divergence dates of between ~3.5 – 4.5 Ga across the gene trees. This places their origin prior to the oldest known fossils, which come from the Strelley Pool Formation (Sugitani et al. 2013; Sugitani, Mimura, Takeuchi, Lepot, et al. 2015; Wacey 2010; Wacey et al. 2011; Javaux 2019). The divergence dates for eukaryotes are some of the most striking results within the gene trees often placing the origin of the group prior to 2 Ga (Fig. 3.18). These dates for eukaryotes contradict the currently prevailing trend of thought for the crown groups origin, which both molecular clock studies (Betts et al. 2018; Eme et al. 2014; Parfrey et al. 2011), as well as fossil evidence (Butterfield 2015a; Knoll and Nowak 2017), place in the Meso- to Paleoproterozoic post 2 billion years. The much older date found here would suggest that eukaryotes are a very ancient group. However, once again, this is most likely due to the smaller amount of information provided by the alignment in the case of a gene by gene basis meaning there is less power to accurately estimate the age of this node. The ages of LBCA, LECA, crown Alphaproteobacteria and crown Cyanobacteria from either side of the duplication in the concatenated analysis overlap. The ages for crown Archaebacteria do not overlap, perhaps suggesting that on one side of the duplication there are a greater number of more quickly evolving genes, this is also where the greater age for the LUCA node is found. As can be seen in Figure 3.20 a large number of nodes in the two main domains, Eubacteria and Archaebacteria, diverge upwards of 2 billion years ago. Although there is a potential fossil record for these lineages, it is by no means certain and so this methodology allows us to come to the conclusion that the crown lineages evolved in the Archaean – Hadean. The results from the concatenated analysis are once again comparable to those produced by the single gene trees.

The novel analyses presented here provide a new way to think about dating the tree of life, especially with regard to the most ancient nodes. The root node in any analysis will always be the most difficult to date and using genes which have a pre-LUCA duplication is a new and informative way to try and

tackle this issue. The utility of gene duplications is something that is only just beginning to be explored in this area. Our results here, both from the cross-bracing of individual genes and cross-calibration of a combined dataset, present a conclusive age for LUCA between 3.89 and 4.51 Ga. The use of duplication events for rooting the tree of life has been used as a technique since the 1980s but here our novel analyses using a concatenated dataset confirms that the exploitation of these extra sources of information can be extremely beneficial to estimating a timescale for the tree of life and most importantly LUCA. Hence, this work adds an important contribution in terms of what can be achieved using this kind of dataset.

Chapter 4

Dating the origin of eukaryotes using the fossilised birth-death process

Author contributions: The ideas for this work were conceived and contributed to by H.C. Betts, J. O'Reilly, P.C.J. Donoghue, T.A. Williams and D. Pisani. H.C. Betts performed all data collection. Analyses were performed by H.C.B. with advice from J. O'Reilly. The following chapter was written and developed by H.C.B. with comments and suggestions by P.C.J.D. and D. P. H.C.B contributed ~90% of the work presented in this chapter.

4.1 Introduction

Eukaryotes are a key lineage in the tree of life, with huge cellular complexity and a diverse range of multicellular forms, they have transformed our planet through ecosystem interactions helping to create the world we know today (Szathmáry and Smith 1995). Despite this, the timing and nature of eukaryote origins is still uncertain (Embley and Martin 2006; Koonin 2010). The last eukaryotic common ancestor (LECA) is thought to have existed sometime in the Proterozoic (Embley and Martin 2006; Eme et al. 2014; Javaux and Lepot 2018; Parfrey et al. 2011). However, the fossil record around this time is poor, therefore little can be ascertained about when LECA evolved and what order its characteristic features were gained in. For example, hotly contested is the timing of the mitochondrial endosymbiosis event in relation to the acquisition of other eukaryotic features with two main competing hypotheses mitochondria-early (Lane and Martin 2010; Martin et al. 2017) versus mitochondria-late (Pittis and Gabaldón 2016). Geological factors have been thought to contribute to their evolution for example the Great Oxidation Event (GOE) at ~2.4-2.1 Ga, and Neoproterozoic Oxidation Event ~0.8-0.5 Ga (Canfield, Poulton, and Narbonne 2007; Gross and Bhattacharya 2010; Knoll and Nowak 2017; Lenton et al. 2014). Here, we estimate divergence times for crown group Eukaryota in order to examine whether the GOE and NOE are temporally linked to the evolution of the crown Eukaryota themselves, or to that of specific eukaryotic lineages. Additionally, we examine the age of crown group eukaryotes in relation to the age of alphaproteobacterial origins (Betts et al. 2018; Shih and Matzke 2013).

Most of the fossil material that we have available is from the eukaryote branch of the tree of life, a huge wealth of information stretching back into the Proterozoic and becoming more fleshed out towards the recent. The earliest possible fossil eukaryotes are mostly acritarchs (Fig. 4.1a-d). These are single celled organic microfossils which, while they most probably belong within Eukaryota, do not usually possess features ascribable to any specific extant eukaryote lineage. They range in form from processed cells (Fig. 4.1a,b) to smooth (Fig. 4.1c) and ornamented (Fig. 4.1d).

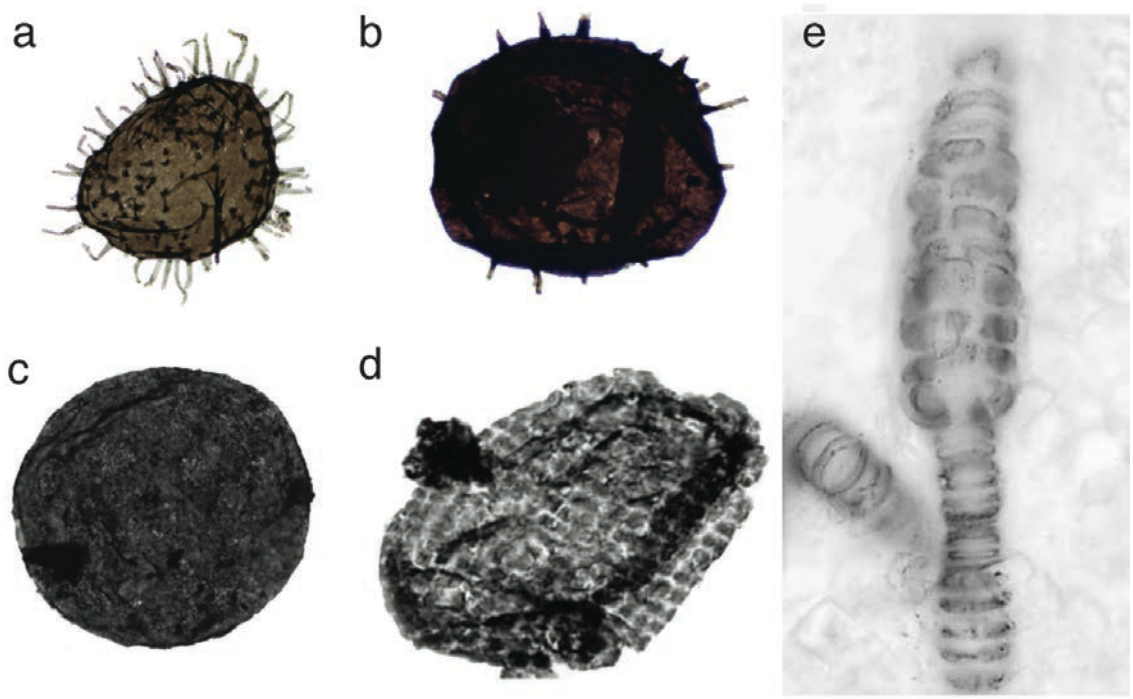


Figure 4.1. A selection of Proterozoic eukaryote fossils. a) acanthomorphic acritarch (Butterfield 2005), b) *Trachyhystrichosphaera aimika* Hermann (Butterfield 2005), c) Smooth walled organic microfossils (Cohen and Macdonald 2015), d) Ornamented organic microfossil (Cohen and Macdonald 2015) and e) *Bangiomorpha pubescens* (Butterfield 2000).

The earliest record of a crown group Eukaryote is that of the Hunting Formation red algal fossil, *Bangiomorpha pubescens* (Fig. 4.1e), which dates to around 1 billion years ago (Butterfield, Knoll, and Swett 1990; Butterfield 2000). It is a multicellular and relatively complex organism which has been identified as a total-group rhodophyte based upon its developmental characters and the distinct shape of its cell arrangements (Betts et al. 2018; Yang et al. 2016). This fossil cannot confidently be resolved as a crown group rhodophyte thus it provides a constraint on the age of the total-group rhodophyte lineage. Older possible red algal fossils date to 1.6 Ga from the Chitrakoot formation (Bengtson, Sallstedt, et al. 2017). However, this fossil is somewhat uncertain in its affinities and it relies upon an accurate interpretation, both of sub-cellular structures as rhomboidal starch granules, and, of gaps between the fossilised cells as pit plugs.

Previously *Bangiomorpha* has been used in node calibrated analyses to find divergence dates for eukaryotes (Betts et al. 2018; Eme et al. 2014; Parfrey et al. 2011). While node calibration can be

extremely useful and provides a clear way in which to date the tree of life, it is limited in terms of the number of fossils that we can apply. Any fossil which cannot be confidently placed within the tree as an oldest member of a specific clade, for instance the acritarch fossils mentioned above, is excluded from the analysis. Thus, we are likely missing the evolution of the early stages of eukaryotes simply due to the poor preservation potential. Furthermore, in node calibration each fossil must be specified using a somewhat arbitrary prior which should be based on a good knowledge of that fossil. Hence, despite the wealth of fossil information early eukaryote origins are still foggy as the record deteriorates as we reach the Proterozoic and node calibration has to exclude the majority of the early eukaryotic fossil record.

To try and tackle the issues associated with node calibration we can instead use the fossilised-birth-death (FBD) process (Heath, Huelsenbeck, and Stadler 2014; Stadler 2010). This model, like the models used by standard molecular clock analyses, takes into account the birth rate, when lineages speciate, and death rate, when lineages go extinct, but also adds on an additional parameter, fossil sampling rate. In so doing it explicitly incorporates fossils into the tree under the same macroevolutionary process. The FBD process allows the analysis to include fossils with uncertain placement and removes the need to assign an arbitrary prior probability distribution to model the age of calibrations. In older time periods, where many fossils cannot be easily assigned to any one scion of life, node calibration is restrictive. In this chapter I have applied the unresolved FBD model to a eukaryotic data set to evaluate whether it is possible to improve our understanding of the timescale of early eukaryote evolution (as detailed in Betts et al. 2018 and Chapter 2).

4.2 Materials and Methods

4.2.1 Molecular dataset

For the initial analysis we used a dataset of 3000 amino acid sites randomly sampled from a concatenated 29 gene dataset. Using the complete alignment meant that it was difficult to achieve convergence. The genes were aligned using MUSCLE (Edgar 2004) before undergoing trimming via TrimAl (Capella-Gutiérrez, Silla-Martínez, and Gabaldón 2009) on the -strict setting. The trimmed gene alignments were then concatenated using FASconCAT (Kück and Meusemann 2010) and it was this final concatenated alignment which was subsampled. The topology for the analysis was fixed as seen in (Betts et al. 2018) and Chapter 2.

4.2.2. Fossil calibrations

In total 56 fossil taxa were used to calibrate the analysis. Six of these are detailed in Chapter 2 as the eukaryotic calibrations. In order to find more fossils for the analysis we used the Palaeobiology Database (PBDB) (Peters and McClennen 2016) which has > 1 million eukaryote records. To this we added a list of Proterozoic eukaryote fossils compiled by Cohen and MacDonald (2015). The latter database contains a comprehensive record of Proterozoic fossils categorised according to their basic morphology. Together these records create an enormous and rich dataset. However, to use all these of these fossils to calibrate the FBD analysis would cause issues of convergence to arise. To allow convergence to be achieved we had to limit subsampling of fossils to 50. The fossils were sampled by reading the complete csv file into R and randomly extracting 50 rows from the dataset (sample_n(df, 50)). In each case the fossil minimum and maximum age was extracted along with its phylogenetic affinity. All fossils were constrained to be members of crown eukaryote groups, aside from *Bangiomorpha pubescens* (Butterfield, Knoll, and Swett 1990; Butterfield 2000) which was set to be a total group rhodophyte. The six fossils from Chapter 2 (Betts et al. 2018) provided minimum calibrations for key nodes in order that these were not under-estimated (O'Reilly and Donoghue 2019) (Fig. 4.2). We applied the unresolved FBD model where no morphological information is available and, therefore, fossil placement is integrated over as part of the mcmc analysis.

4.2.3 Divergence time analyses

Molecular dating was carried out using BEASTv2 (Bouckaert et al. 2014); accounting for sampled ancestors (Gavryushkina et al. 2014), with the uncorrelated lognormal (UCLN) relaxed clock and the LG model of amino acid substitution. The origin of the process was assigned a prior stretching between the maximum age of the oldest included fossil, *Bangiomorpha pubescens* 1279 Ma, and the age of the Earth 4520 Ma (Barboni et al. 2017; Hanan and Tilton 1987). Rho, the parameter that represents the percentage of extant lineages from the group of interest included in the analysis, was fixed to the proportion of extant eukaryote species ($2.65x10^{-6}$) based upon an estimate of present species numbers from (Mora et al. 2011).

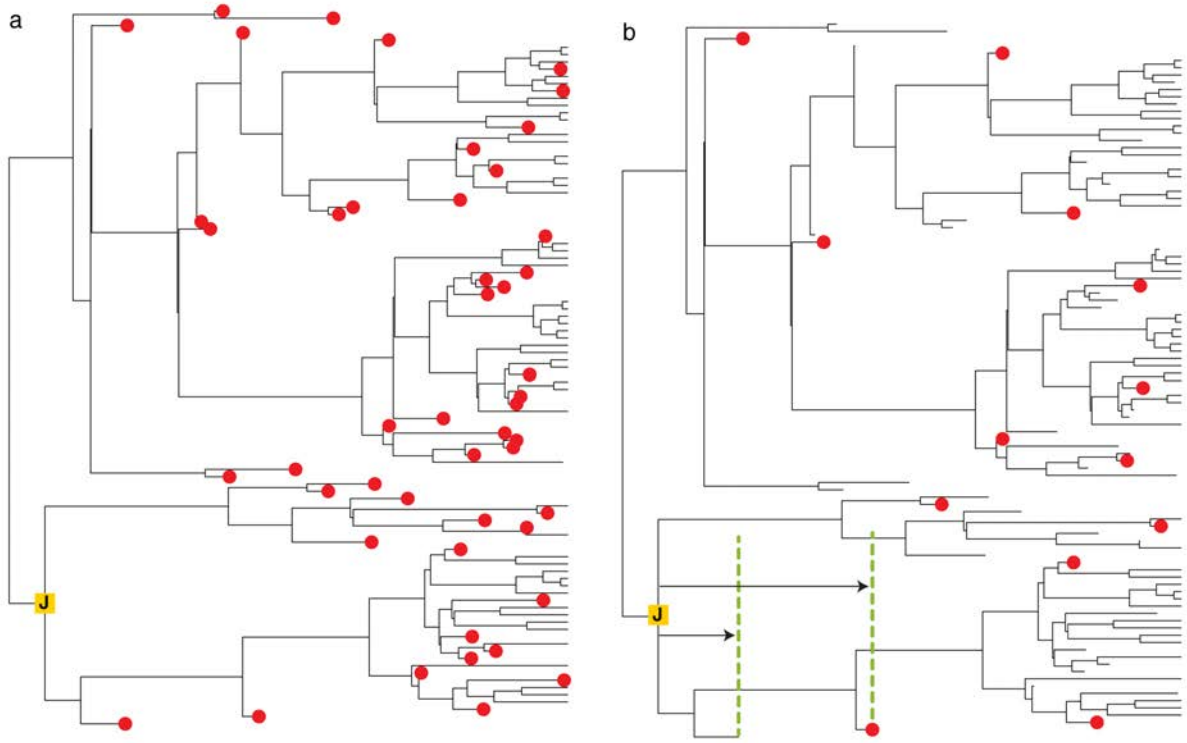


Figure 4.2. a) A tree with a complete sample of fossil calibrations – shown as red dots, b) If we take a completely random subsample of the fossils, we can see that node J in yellow has the risk of being underestimated. This is because the oldest possible calibrating fossil is not sampled.

The FBD skyline model in the bdsky package (Stadler et al. 2013) was applied to all our analyses. This splits the process into distinct time slices from the origin of the process to the present. Within the time slices parameter values remain consistent. Between them they can be set to vary. This allows for more

specific modelling between different time intervals, though does not account for differences between lineages. Allowing changes between time slices can better model some events if something specific is known about the evolution of a group or major geological changes which may have affected it. Here, the analysis was split into 3 time slices which were allocated as 4520 – 1000, 1000 – 539.2 and 539.2 – 0 Ma. The parameter that we looked at with reference to this was the rate of fossil sampling. In our initial analyses the rate of fossil sampling was set to be constant between time slices. However, we know that the quality of the fossil record deteriorates back in time, dropping off dramatically into the Precambrian. For this reason, we undertook sensitivity analyses using the rate of fossil sampling which are detailed in the next paragraph, allowing the rate to decrease in older time slices.

4.2.3.1 Divergence time sensitivity analyses

Sensitivity analyses were performed, to check whether any of the parameters were having an effect on the divergence time, as follows. Firstly, a different subsample of 3000 sites from the amino acid alignment was used, thus we used two subsamples. Secondly, an older possible rhodophyte fossil was used, *Rafatazmia chitrakootensis* (Bengtson, Sallstedt, et al. 2017), in place of *Bangiomorpha*. This was carried out in order to test the robustness of the analysis to a marked change in calibration. Additional analyses in the same category were performed with the inclusion of 10 fossils sampled from the Cohen and MacDonald (2015) database of eukaryotic fossils. This is an alternative to PDBD that is more complete. In one analysis we specifically sampled 10 of these fossils and added them to the dataset, resulting in 66 fossil calibrations overall. This was done because they were not sampled as part of the random 50 from the combined PBDB + Cohen and MacDonald dataset and we wanted to examine whether including a larger sample, of more ancient fossils, would influence the outcome. Fossils from the Cohen and MacDonald database were assigned a total group eukaryote affinity. In two more analyses the rho (extant sampling) parameter was varied, firstly to 0.1 and then to 0.001. Finally, the fossil sampling rate was varied in the different time bins in two ways. Firstly, arbitrarily with a value of 0.001, 0.01 and 0.1 in each of the three time-bins from the oldest to the most recent. Secondly, with maximum likelihood estimate (MLE) rates of fossil sampling following the (Solow and Smith 1997) method implemented using the R package PaleoTree R (Bapst et al. 2016) using the full sample of

eukaryote data from which we sampled taxa for our analysis. For every single analysis five separate chains were run to convergence which was assessed using the package Tracer. In each case all runs from the analysis which reached convergence from were combined to produce the final divergence times and fossils were cut from the tree before being summarised.

4.3 Results

Results from the initial analysis (*Bangiomorpha* oldest fossil, rho value = $2.65x10^{-6}$ and rate of fossil sampling = 0.1 in all time slices) suggest that the LECA evolved between 1038-1077 million years ago (Ma) in the late Mesoproterozoic (Fig 4.3a). The stem lineage divergences of Excavata, SAR (stramenopiles, alveolates, rhizarians) and Archaeplastida occur in quick succession in the next 8 million years just prior to the age of *Bangiomorpha* (Fig. 4.3b). There is a time lag to the divergence of the crown lineages, for example, Metazoa between 551-563 Ma (Fig. 4.3c), fungi between 394-435 Ma (Fig. 4.3d) and SAR between 535 – 586 Ma (Fig. 4.3e). These divergence time dates remain consistent if we use a different 3000 site amino acid sample taken from the concatenated alignment. This is shown by the similarity between the divergence dates for LECA (Fig. 4.4) across 5 different amino acid samples used. When *Rafatazmia chitrakootensis* is used as the oldest crown group fossil, ~1500 Ma when compared to *Bangiomorpha's* ~1000 Ma, there is an increase in the divergence age of crown eukaryotes consistent with this jump in time (Fig. 4.3a) to between 1568-1594 Ma in the early Mesoproterozoic. Despite this jump in time the credible intervals are still very narrow across the tree (Fig. 4.3).

The ages of the crown lineages as mentioned above are; SAR 533-579 Ma, green plants 450-497 Ma, Archaeplastida 1561-1568 Ma and Metazoa 550-563 Ma. Hence, the only nodes affected by the use of the Chitrakoot fossil are those that fall above the stem rhodophytes (Fig. 4.3 a,b) and any nodes towards the tips retain the same divergence dates as when *Bangiomorpha* is the oldest calibrating fossil (Fig. 4.3c-e). If we keep *Bangiomorpha* as the oldest fossil and include 10 fossils from the Cohen and MacDonald 2015 acritarch dataset very similar divergence dates are produced once again. The age of LECA does not change (Fig. 4.3a) when compared with analyses that do not include these ancient acritarch fossils, and the fossils themselves are placed on the stem branch leading to LECA. There is also no effect on nodes further down the tree (Fig. 4.3b-e). If the extant sampling is altered to be less representative of the likely sampling proportion (rho = 0.1 and 0.001 instead of $2.65x10^{-6}$) the resulting

divergence dates are similar to the standard analysis, most often only being 1-2 million years difference in minimum and maximum value for the crown nodes (Fig. 4.3).

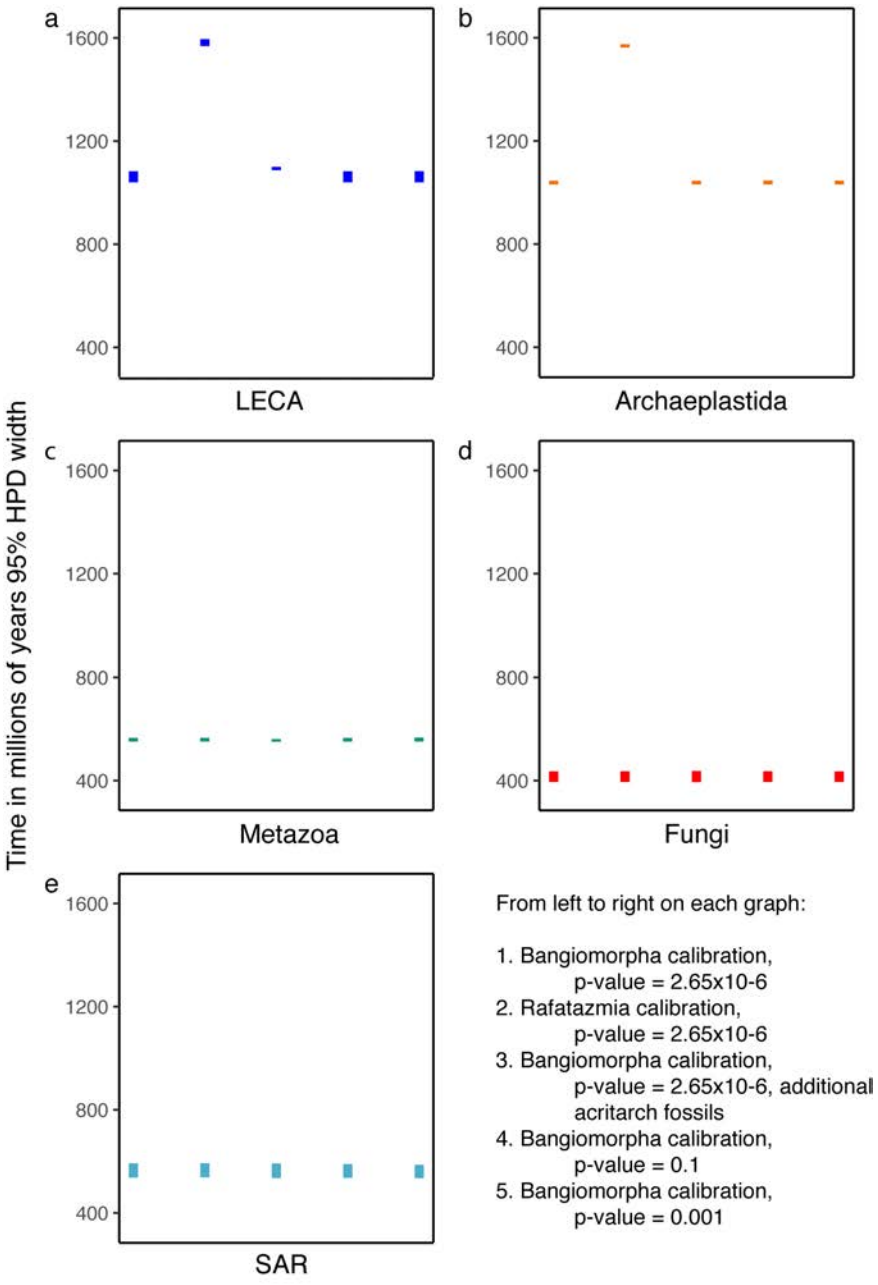


Figure 4.3. Examples of the credible intervals for 5 nodes within the tree under 5 different analysis set ups using the FBD; a) the last eukaryotic common ancestor, b) Archaeplastida, c) Metazoa, d) fungi and e) the stramenopiles, alveolates and rhizarians (SAR.)

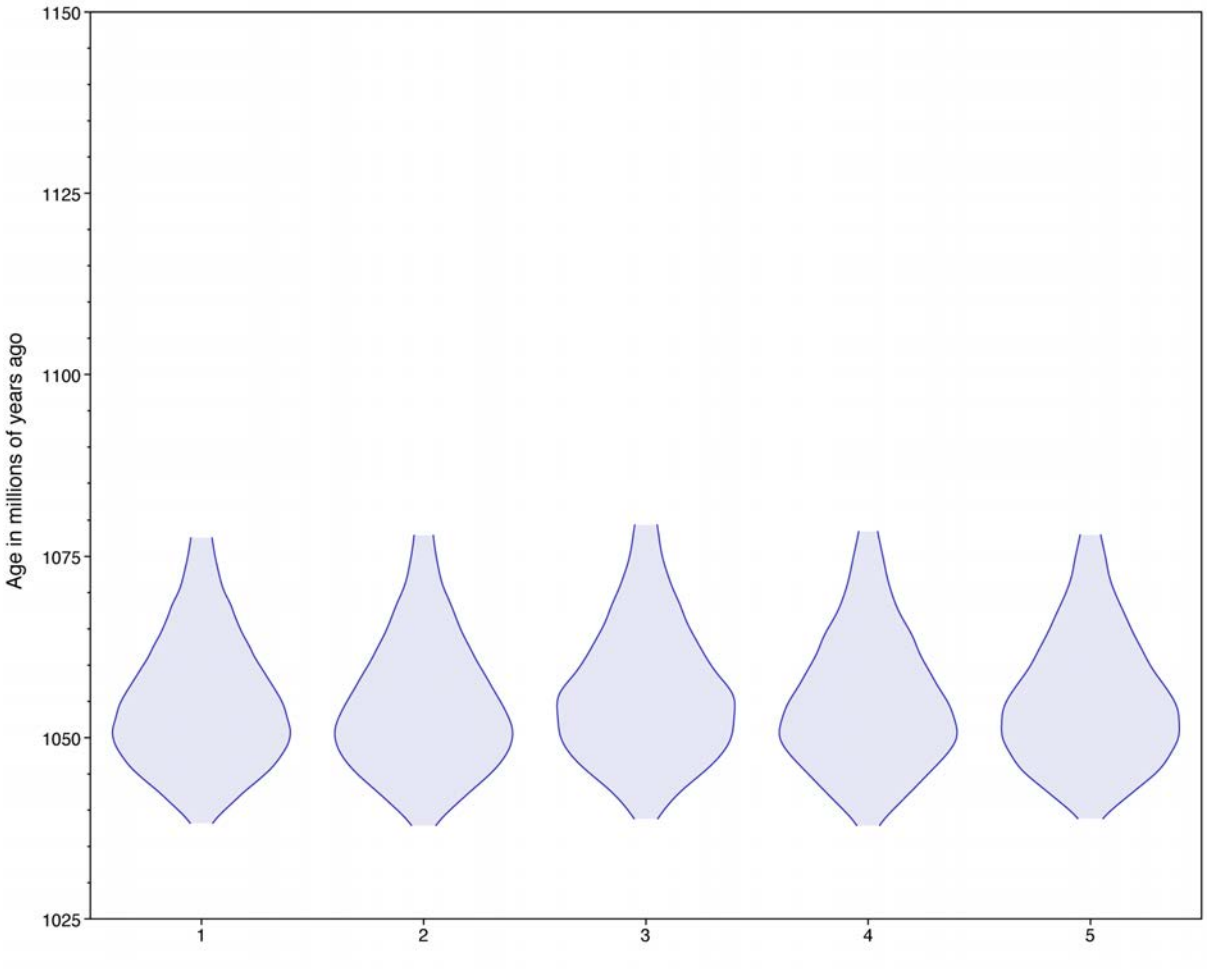


Figure 4.4. The divergence times (95% HPD) for the last eukaryotic common ancestor for analysis carried out using 5 different amino acid samples of 3000 sites. Each sample was taken from the same concatenated alignment.

The most distinctive difference in divergence dates is produced when the rate of fossil sampling is allowed to vary between different time slices via use of the fossilised birth-death skyline model. When we specify arbitrary rates of 0.001, 0.01 and 0.1 for the oldest to youngest time slices respectively the divergence dates increase in all parts of the tree. Crown eukaryotes split at 1125-1647 Ma, crown Archaeplastida at 1032-1240 Ma and crown Metazoa between 550.7-658 Ma (Fig. 4.5). The oldest nodes have much larger credible interval widths than seen in the other analyses, spanning hundreds of millions of years rather than only tens or even less than this. The wide credibility intervals are further exacerbated by the implementation of the ML estimated rates of fossil sampling using (Solow and Smith 1997) with LECA being proposed to have a divergence date somewhere between 1247 – 1884 Ma (Fig. 4.5).

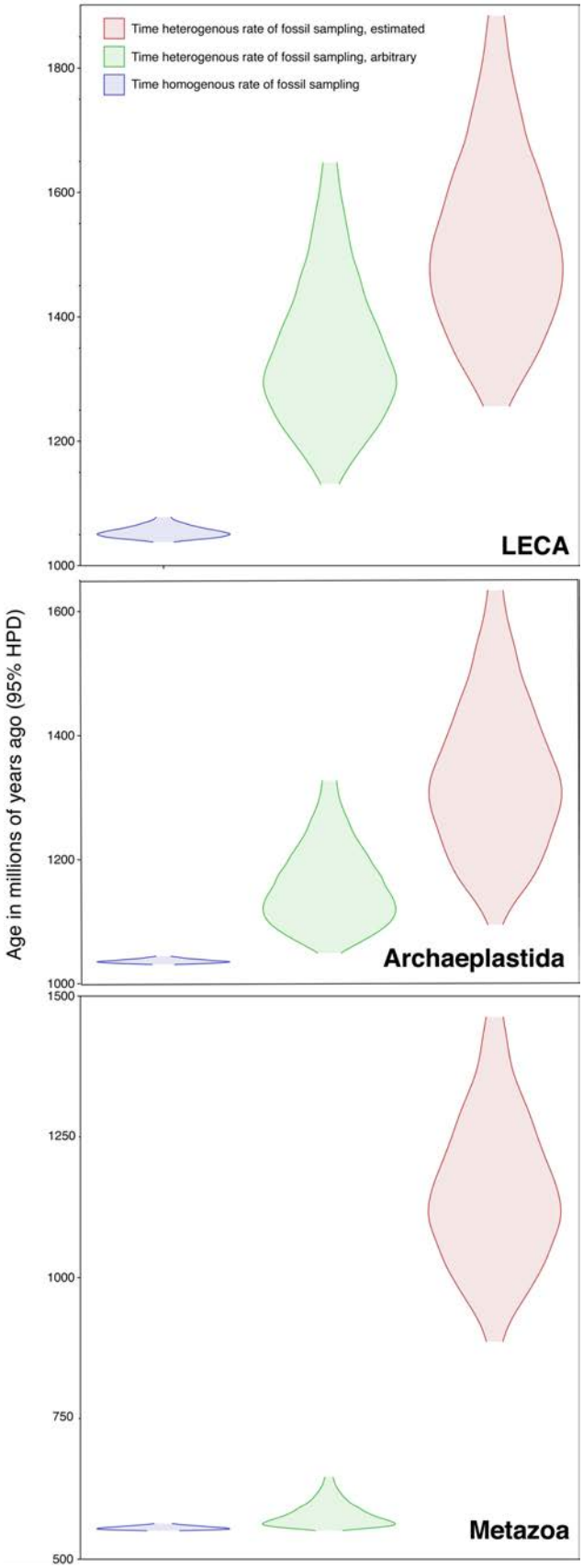


Figure 4.5. Estimates of divergence time for 3 key nodes in the eukaryote tree; the last universal common ancestor, Archaeplastida and Metazoa. In each case the effect of three different approaches to modelling the rate of fossil sampling using the skyline model can be seen.

4.4 Discussion

The results from varying the input and parameters show a trade-off between volume of information and computational power. In order to achieve acceptable mcmc sampling for the majority of parameters it was necessary to use a small number of extant taxa as well as a small number of amino acid sites. Although, changing the subsample of amino acids sites did not have an appreciable effect upon the output (Fig. 4.4). The number of fossils included is much less than the total possible sample which numbers millions. Adding either additional fossil records or extant taxa would cause the mcmc sampling to be poor making it difficult to reach convergence. This implies that at the moment it is too computationally expensive to use the method to its full potential. The 95% HPD widths in most cases across the analyses presented here are very small (Fig 4.3, Fig. 4.6a). While this high precision could be useful when looking at the co-evolution of eukaryotes and the geological changes around them, it is worrying that these results might be reminiscent of false precision, and that we are therefore not capturing the real divergence time estimates. The risk of underestimating the ages of nodes applying the FBD is still prevalent, something which seems less likely in node calibration These results suggest that the FBD should be used with caution over very long timescales, with the correct application of the rate of fossil sampling and thought into how many fossils are appropriate to add into the analysis. It seems that because of these nuances, and the lack of genetic data that can be included, at such timescales the FBD does not make a superior replacement for node calibration practices.

The results of the various sensitivity analyses overlap but do not present a homogenous timeline for the evolution of eukaryotes. The youngest minimum age for LECA is recovered as 1038 Ma and the greatest maximum age is recovered as 1884 Ma. This broad spread of dates reflects the findings of previous studies (Betts et al. 2018; Eme et al. 2014; Parfrey et al. 2011). In cases where the rate of fossil sampling is homogeneous across time, regardless of whether other parameters in the analysis are varied, the age of LECA is close to the age of the oldest calibrating fossil (Fig. 4.3). Additionally, for each node the 95% credible interval very small, this can be clearly seen in Figure 4.6a where the divergence dates form a cluster right next to the age of *Bangiomorpha*. Hence, when *Bangiomorpha* is the oldest fossil

(1030 Ma) we find a date for LECA of between 1038-1077 Ma and when *Rafatazmia* is the oldest fossil (1561 Ma) we find a date of between 1568-1594 Ma. These two estimates have incredibly narrow credible intervals, neither of which seem likely to encompass the true divergence date, a case of precision over accuracy. Whereas, when the rate of fossil sampling is allowed to vary either arbitrarily (Fig. 4.6b), or using estimates (Fig. 4.6c), much more realistic divergence dates for LECA are produced (1125-1647 Ma and 1247-1884 Ma respectively).

The issues of small credible intervals in analyses with a time homogenous rate of fossil sampling are reflected on the more tip-wards divergences, with nodes often falling just prior to any fossil calibrations. In cases examined here, any when the rate of fossil sampling is not estimated, the age of crown Metazoa is extremely young (between ~551 – 650 Ma), clustered at the minimum date (Fig 4.3 and 4.5). This is young when we consider potential fossil records of metazoan taxa, as well as results from other molecular clock analyses. The oldest confirmed metazoan fossil is usually thought to be *Kimberella quadrata* which has an age of just over 550.25 Ma (Narbonne et al. 2012). However, other potential Metazoan fossils date to around this time as well as the Ediacaran fauna which some researchers believe to be metazoan in nature (Dunn, Liu, and Donoghue 2018; Hoekzema et al. 2017). These potential metazoan fossils are more ancient than *Kimberella*. The divergence dates are also much younger than other molecular clock estimates which place the origin of this group between ~850-650 Ma (Erwin et al. 2011; dos Reis, Donoghue, and Yang 2015; Betts et al. 2018; Peterson and Butterfield 2005). These results are produced despite a much larger inclusion of fossil taxa than has previously been used in molecular dating estimates and is supposedly the great strength of the FBD analyses. However, when the rate of fossil sampling throughout is allowed to vary time wider credible intervals (1261 – 1901 Ma) are produced.

Despite concerns laid out above when appropriately applied the FBD skyline model does seem to produce reasonable results (Fig. 4.6c) and shows good agreement with previous divergence date estimates of LECA (Betts et al. 2018; Eme et al. 2014; Parfrey et al. 2011). This meant using the oldest fossil we were confident of, *Bangiomorpha pubescens* and allowing the rate of fossil sampling to vary

between the time slices using estimated values. Some estimates have placed LECA back into the Palaeoproterozoic far enough to coincide with the GOE (Hedges et al. 2004) an event which is considered to potentially have had an effect on eukaryote evolution (Knoll and Nowak 2017). However, the emerging pattern (Betts et al. 2018; Eme et al. 2014; Parfrey et al. 2011), backed up by these results, is that LECA was very much a post-GOE organism (Fig. 4.5, Fig. 4.6), that may have benefitted from an oxygenated world, but did not seem to be initially stimulated by it. This is not to say that the transition from the first eukaryotic common ancestor (FECA) to LECA was immune to such environmental upheaval, only that the radiation of the crown group, sometimes speculated to be rapid as well as influenced by increased oxygen, was not. This tree (4.6c) also exhibits older, though perhaps more agreeable results for the metazoans. Though at between 1071–1686 Ma this does push the origin of metazoans back into the Mesoproterozoic, prior to the Neoproterozoic oxygenation event. This could back up suggestions that metazoans had a hand in stimulating the event, via an increase in benthic filter feeding, as well as the sinking of eukaryotic particles shifting the consumption of oxygen away from surface waters (Lenton et al. 2014). The evolution of crown Archaeplastida also occurs around this time (1108-1671 Ma). This coincides with previous estimates (Betts et al. 2018; Eme et al. 2014; Parfrey et al. 2011; Sánchez-Baracaldo et al. 2017; Brocks et al. 2017). However, it is still a long time before the suggested rise to ecological dominance of the group based upon biomarker evidence (Brocks et al. 2017).

In most cases presented above the analyses are either being driven by the fossil calibrations, or by the rate of fossil sampling parameter. In either case, the limitations on the number of fossils that can be included, and the amount of molecular data that we can practically include, shows that including the complete amount of fossil data is not currently feasible for such long timescales and for so much fossil data. Additionally, because we are applying the same fossils as in a node-calibration, taken from the Betts et al., 2018 analysis, this means we have not moved away from a fundamental oldest-fossil method. However, on the whole results presented here suggest that LECA was a post-GOE organism and that the crown members of many eukaryote clades emerged during the boring billion (1.8-0.8 Ga).

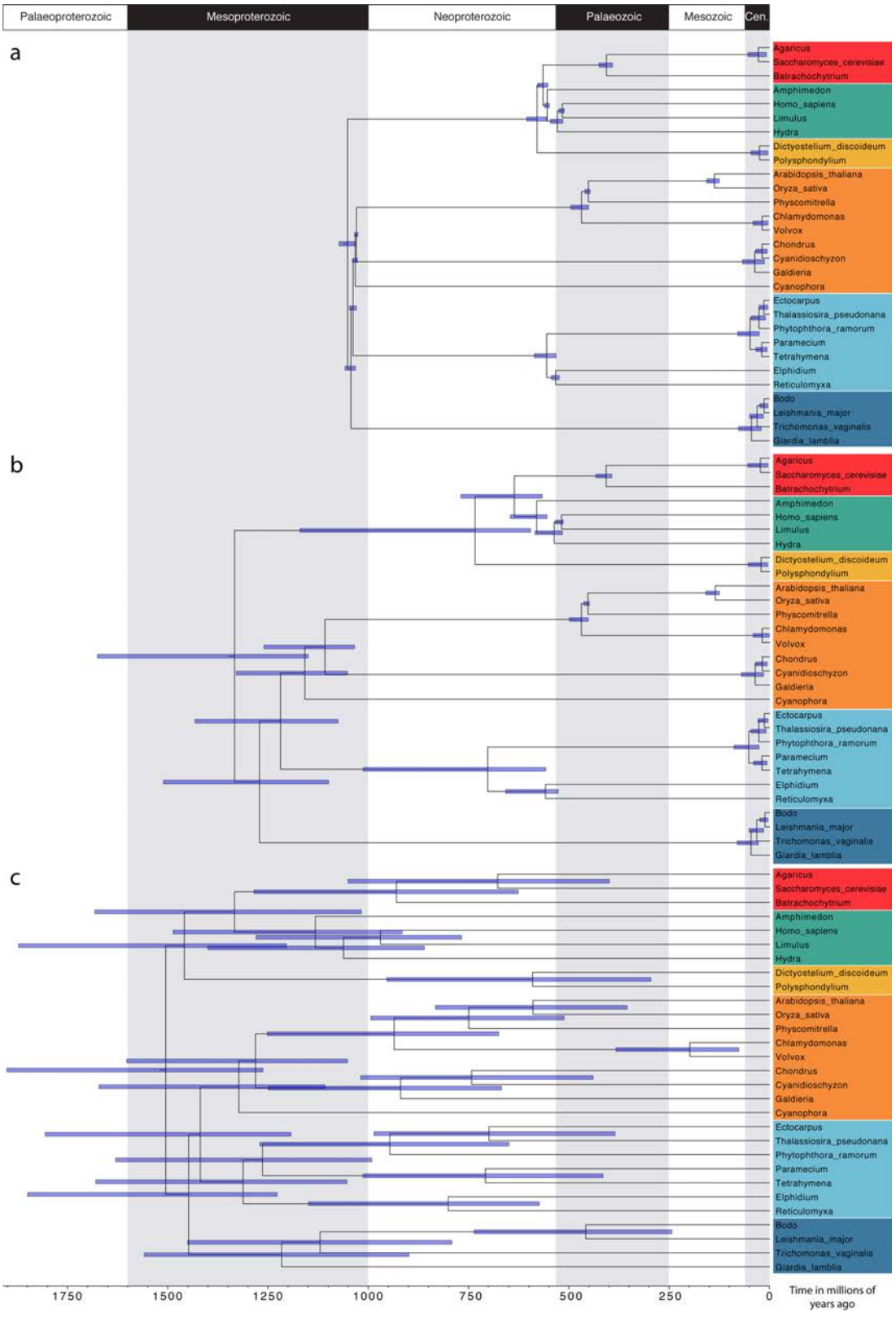


Figure 4.6. Divergence time trees for three different analyses; a) an analysis carried out with a homogenous rate of fossil sampling in all time slices, b) an analysis with a heterogenous rate of fossil sampling with arbitrary values and c) an analysis with a heterogenous rate of fossil sampling with estimated values. In each case the colours represent different clades. Red = fungi, green = metazoans, yellow = ameobozoans, orange = Archaeplastida, light blue = SAR and dark blue = excavates. The pale blue bars indicate the 95% credible interval.

Chapter 5

Conclusions

Author contributions: This chapter was written and developed by H.C. Betts. H.C.B. contributed 100% of the work presented in this chapter.

Tracking the early history of life and indeed the early history of key lineages within life has traditionally been approached using the fossil record. However, this approach, while it is useful, is unsuitable for dating the origin of any lineage as fossils belonging to that lineage take time to accumulate the features necessary for its identification within the record. This is especially true of those lineages which have poor fossilisation potentials, arose a long time ago, meaning there is a lack of material, or which have few distinctive morphological features (Javaux 2019). Consequently, associating them with an extant lineage is not possible. This thesis has attempted to update the understanding of how best to estimate divergence times for anciently diverging nodes within the tree of life, looking at the challenges of producing robust fossil calibrations, and what novel approaches can be used to tackle such problems. All of which has made use of molecular clock methodology. Both well establish node calibration methods, applying these calibrations using cross-bracing, as well as making use of the fossilised birth death-process.

Throughout this thesis some key nodes have been of greater interest. These are the last universal common ancestor (LUCA), the last bacterial common ancestor (LBCA), the last archaeal common ancestor (LACA) and the last eukaryotic common ancestor (LECA). These are all noteworthy because they mark the appearance of fundamental body plans and sets of characteristics which still inform the domains as we know them today. The analyses presented here suggest that LUCA was a very ancient organism that existed close to the formation of the planet (> 4.5 Ga). Crucially, even though around this time there is evidence for major planetary systems emerging (Harrison, Bell, and Boehnke 2017), it is still prior to the end of the Late Heavy Bombardment. This suggests that early life possibly existed in small resistant habitats and that it must have been tough enough to survive the huge upheaval undergone by the early Earth. This age for LUCA, first found in our combined analysis (Chapter 2), is confirmed by the analysis both of individual genes and also of a concatenated duplicated dataset (Chapter 3). The addition of a node before LUCA helped to provide more information for estimating its divergence times. Hence, allowing us to be more confident of our predicted age for LUCA which falls between ~4-4.5 Ga. This ancient date is also reflected in the divergence times for LBCA and LACA which both appear in a relatively similar timeframe between ~3.5 – 4.0 Ga. This is just prior to the time that confirmed

fossils begin to appear in the record, for example those from the Strelley Pool formation (Sugitani et al. 2013; Sugitani, Mimura, Takeuchi, Lepot, et al. 2015; Wacey 2010; Wacey et al. 2011; Javaux 2019).

As with LUCA our estimates for LECA, produced using different analysis methods (concatenation and the fossilised birth death process), are roughly congruent. Currently LECA is still a widely debated organism, in terms of timing and characteristics, and our results add to this debate by suggesting that the eukaryotic common ancestor evolved within the Palaeo-Meso-Proterozoic. This time is suggested to extend as back as the Archaean in Chapter 3. However, this is likely due to the information available in individual gene trees and their resultant large credibility intervals. The main radiation of eukaryotes occurs during the boring billion in both Chapters where eukaryotes are a key focus (Chapter 2 and 4). This is a period of little geological change and supposedly little biological change. However, the results presented within this thesis suggest that major radiations of key eukaryotic lineages were occurring during this time. It is also far after the great oxidation event which may have produced the oxygen necessary for the radiation of eukaryotes. However, they did not seem to be initially stimulated by it despite the fact that the first eukaryotic common ancestor may have benefitted from this environmental upheaval. In our analyses the ages of the alphaproteobacterial crown and the cyanobacterial crown groups also align with their endosymbiotic gene transfer groups. In the case of the mitochondrial endosymbiosis event this offers tantalising evidence that this major evolutionary transition happened in close association with the evolution of the eukaryote crown group, perhaps even stimulating its emergence. Though this cannot unfortunately help to answer the question of whether this event was early or late in terms of how formed the ancestral eukaryotic cell was (Martin et al. 2017; Pittis and Gabaldón 2016).

This thesis has pushed the boundaries of divergence time estimation methods in order to try and elucidate a timescale for the whole of life. This involved the employment of up to date and new methodology, especially the application of cross-bracing and cross-calibration which has so far not been applied to such an ancient divergence using a concatenated dataset. Here it has been used with ancient gene duplication events to help pin down the divergence date of the last universal common ancestor.

Additionally, the fossilised birth death process has been employed an attempt to harness the information available in our rich and large fossil record. When employed with all parameters taken properly into account this methodology can provide a useful way to incorporate more fossil material.

Future directions for this research could lie in further harnessing of these new methodologies. especially when the software for cross-bracing allows this to be used on a concatenated duplicated tree. In addition to this more could be done to investigate the endosymbiotic gene transfer events, they have been woven throughout this thesis, but their further investigation would be of great benefit to our understanding of the evolution of life. I feel confident that this thesis has added to the scope and breadth of knowledge concerning a timeline for life on our planet. The integrative approach of using fossil data with the molecular clock has helped to produce a robust timescale for the tree of. Hopefully, it will help to contribute to debates about when an ancestor for life might have existed and therefore in what conditions it might have emerged and what upheaval it may have survived. As the timeline for life becomes more and more refined so improved inferences about the influence of geochronological factors on evolutionary process can be made.

References

Aberer, A. J., Krompass, D. and Stamatakis, A. 2012. 'Pruning Rogue Taxa Improves Phylogenetic Accuracy: An Efficient Algorithm and Webservice', *Systematic Biology*, 62: 162-66.

Abhishek, A., Bavishi, A., Bavishi, A., and Choudhary, M. 2011. 'Bacterial genome chimaerism and the origin of mitochondria', *Canadian Journal of Microbiology*, 57: 49-61.

Abramov, O., and Mojzsis, S. J. 2009. 'Microbial habitability of the Hadean Earth during the late heavy bombardment', *Nature*, 459: 419.

Allwood, A. C., Walter, M. R., Burch, I. W., and Kamber, B. S. 2007. '3.43 billion-year-old stromatolite reef from the Pilbara Craton of Western Australia: Ecosystem-scale insights to early life on Earth', *Precambrian Research*, 158: 198-227.

Allwood, A. C., Walter, M. R., Kamber, B. S., Marshall, C. P., and Burch, I. W. 2006. 'Stromatolite reef from the Early Archaean era of Australia', *Nature*, 441: 714-18.

Altermann, W., and Schopf, J. W. 1995. 'Microfossils from the Neoarchean Campbell Group, Griqualand West Sequence of the Transvaal Supergroup, and their paleoenvironmental and evolutionary implications', *Precambrian Research*, 75: 65-90.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. 1990. 'Basic local alignment search tool', *Journal of Molecular Biology*, 215: 403-10.

Amard, B., and Bertrand-Sarfati, J. 1997. 'Microfossils in 2000 Ma old cherty stromatolites of the Franceville Group, Gabon', *Precambrian Research*, 81: 197-221.

Anbar, A. D. et al. 2007. 'A whiff of oxygen before the great oxidation event?', *Science*, 317: 1903-06.

Antcliffe, J. B., Gooday, A. J., and Brasier, M. D. 2011. 'Testing the protozoan hypothesis for Ediacaran fossils: a developmental analysis of Palaeopascichnus', *Palaeontology*, 54: 1157-75.

Atteia, A. et al. 2009. 'A Proteomic Survey of Chlamydomonas reinhardtii Mitochondria Sheds New Light on the Metabolic Plasticity of the Organelle and on the Nature of the α-Proteobacterial Mitochondrial Ancestor', *Molecular biology and evolution*, 26: 1533-48.

Baldauf, S. L., Palmer, J. D., and Doolittle, W. F. 1996. 'The root of the universal tree and the origin of eukaryotes based on elongation factor phylogeny', *Proceedings of the National Academy of Sciences*, 93: 7749-54.

Bapst, D. W., Wright, A. M., Matzke, N. J., and Lloyd, G. T. 2016. 'Topology, divergence dates, and macroevolutionary inferences vary between different tip-dating approaches applied to fossil theropods (Dinosauria)', *Biology Letters*, 12: 20160237.

Barboni, M. et al. 2017. 'Early formation of the Moon 4.51 billion years ago', *Science Advances*, 3: e1602365.

Baross, J. A., and Hoffman, S. E. 1985. 'Submarine hydrothermal vents and associated gradient environments as sites for the origin and evolution of life', *Origins of Life and Evolution of the Biosphere*, 15: 327-45.

Bekker, A. et al. 2004. 'Dating the rise of atmospheric oxygen', *Nature*, 427: 117-20.

Bengtson, S. et al. 2017. 'Fungus-like mycelial fossils in 2.4-billion-year-old vesicular basalt', *Nature Ecology &Amp; Evolution*, 1: 0141.

Bengtson, S., Sallstedt, T., Belivanova, V., and Whitehouse, M. 2017. 'Three-dimensional preservation of cellular and subcellular structures suggests 1.6 billion-year-old crown-group red algae', *PLOS Biology*, 15: e2000735.

Benson, D. et al. 2013. 'GenBank', *Nucleic Acids Research*, 42: D32-D37.

Benton, M. J. et al. 2015. 'Constraints on the timescale of animal evolutionary history', *Palaeontologia Electronica*, 18: 1-106.

Benton, M. J., and Donoghue, P. C. J. 2007. ‘Palaeontological evidence to date the tree of life’, *Molecular Biology and Evolution*, 24: 26-53

Benton, M. J., Ruta, M., Dunhill, A. M., and Sakamoto, M. 2013. 'The first half of tetrapod evolution, sampling proxies, and fossil record quality', *Palaeogeography, Palaeoclimatology, Palaeoecology*, 372: 18-41.

Betts, H.C. et al. 'Integrated genomic and fossil evidence illuminates life's early evolution and eukaryote origin', *Nature Ecology & Evolution*, 2: 1556-62.

Bonen, L., Cunningham, R. S., Gray, M. W., and Doolittle, W. F. 1977. 'Wheat embryo mitochondrial 18S ribosomal RNA: evidence for its prokaryotic nature', *Nucleic Acids Research*, 4: 663-71.

Borrel, G., Panagiotis S. A., and Gribaldo., S. 2016. 'Methanogenesis and the Wood–Ljungdahl Pathway: An Ancient, Versatile, and Fragile Association', *Genome biology and evolution*, 8: 1706-11.

Bosak, T., Macdonald, F., Lahr, D. and Matys, E. 2011. 'Putative cryogenian ciliates from Mongolia', *Geology*, 39: 1123-26.

Bosak, T. et al. 2011. 'Agglutinated tests in post-Sturtian cap carbonates of Namibia and Mongolia', *Earth and Planetary Science Letters*, 308: 29-40.

Bosak, T. et al. 2012. 'Possible early foraminiferans in post-Sturtian (716−635 Ma) cap carbonates', *Geology*, 40: 67-70.

Bottke, W. F. et al. 2015. 'Dating the Moon-forming impact event with asteroidal meteorites', *Science*, 348: 321-23.

Bouckaert, R. et al.. 2014. 'BEAST 2: A Software Platform for Bayesian Evolutionary Analysis', *PLOS Computational Biology*, 10: e1003537.

Brasier, M. D. et al. 2002. 'Questioning the evidence for Earth's oldest fossils', *Nature*, 416: 76-81.

Brasier, M. D. et al. 2005. 'Critical testing of Earth's oldest putative fossil assemblage from the ~3.5Ga Apex chert, Chinaman Creek, Western Australia', *Precambrian Research*, 140: 55-102.

Brasier, M. D., McLoughlin, N., Green, O., and Wacey, D. 2006. 'A fresh look at the fossil evidence for early Archaean cellular life', *Philosophical Transactions of the Royal Society B: Biological Sciences*, 361: 887-902.

Brasier, M. D. and Lindsay, J. F. 1998 'A billion years of environmental stability and the emergence of eukaryotes: new data from northern Australia' *Geology*, 26: 555-558.

Brocks, J. J. et al. 2017. 'The rise of algae in Cryogenian oceans and the emergence of animals', *Nature*, 548: 578.

Brown, J. R., and Doolittle, W. F. 1995. 'Root of the universal tree of life based on ancient aminoacyl-tRNA synthetase gene duplications', *Proceedings of the National Academy of Sciences*, 92: 2441-45.

Brown, J. R., Robb, F. T., Weiss, R., and Doolittle, W. F. 1997. 'Evidence for the early divergence of tryptophanyl-and tyrosyl-tRNA synthetases', *Journal of Molecular Evolution*, 45: 9-16.

Buick, R. 1990. 'Microfossil Recognition in Archean Rocks: An Appraisal of Spheroids and Filaments from a 3500 M.Y. Old Chert-Barite Unit at North Pole, Western Australia', *PALAIOS*, 5: 441-59.

Butterfield, N. J. 2000. 'Bangiomorpha pubescens n. gen., n. sp.: implications for the evolution of sex, multicellularity, and the Mesoproterozoic/Neoproterozoic radiation of eukaryotes', *Paleobiology*, 26: 386-404.

Butterfield, N. J. 2004. 'A vaucheriacean alga from the middle Neoproterozoic of Spitsbergen: implications for the evolution of Proterozoic eukaryotes and the Cambrian explosion', *Paleobiology*, 30: 231-52.

Butterfield, N. J. 2005. 'Probable Proterozoic fungi', *Paleobiology*, 31: 165-82.

Butterfield, N. J. 2015a. 'Early evolution of the Eukaryota', *Palaeontology*, 58: 5-17.

Butterfield, N. J. 2015b. 'Proterozoic photosynthesis – a critical review', *Palaeontology*, 58: 953-72.

Butterfield, N. J., Knoll, A. H., and Swett, K. 1994. 'Paleobiology of the Neoproterozoic Svanbergfjellet Formation, Spitsbergen', *Lethaia*, 27: 76-76.

Butterfield, N. J., Knoll, A. H., and Swett, K. 1990. 'A bangiophyte red alga from the Proterozoic of arctic Canada', *Science*, 250: 104-07.

Byerly, G. R., Kröner, A., Lowe, D. R., Todt, W., and Walsh, M. M. 1996. 'Prolonged magmatism and time constraints for sediment deposition in the early Archean Barberton greenstone belt: evidence from the Upper Onverwacht and Fig Tree groups', *Precambrian Research*, 78: 125-38.

Byerly, Gary R., Donald R. Lower, and Maud M. Walsh. 1986. 'Stromatolites from the 3,300–3,500-Myr Swaziland Supergroup, Barberton Mountain Land, South Africa', *Nature*, 319: 489-91.

Cairns-Smith, A. G. 1978. 'Precambrian solution photochemistry, inverse segregation, and banded iron formations', *Nature*, 276: 807-08.

Canfield, D. E., Poulton, S. W., and Narbonne, G. M. 2007. 'Late-Neoproterozoic Deep-Ocean Oxygenation and the Rise of Animal Life', *Science*, 315: 92-95.

Capella-Gutiérrez, S., Silla-Martínez, J. M., and Gabaldón, T. 2009. 'trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses', *Bioinformatics*, 25: 1972-73.

Carlson, R. W., and Lugmair, G. W. 1988. 'The age of ferroan anorthosite 60025: oldest crust on a young Moon?', *Earth and Planetary Science Letters*, 90: 119-30.

Chapman, C. R., Cohen, B. A., and Grinspoon, D. H. 2007. 'What are the real constraints on the existence and magnitude of the late heavy bombardment?', *Icarus*, 189: 233-45.

Charlebois, R. L., Sensen, C. W., Doolittle, W. F., and Brown, J. R. 1997. 'Evolutionary analysis of the hisCGABdFDEHI gene cluster from the archaeon Sulfolobus solfataricus P2', *Journal of bacteriology*, 179: 4429.

Chernikova, D., Motamedi, S., Csürös, M., Koonin, E. V., and Rogozin, I. B. 2011. 'A late origin of the extant eukaryotic diversity: divergence time estimates using rare genomic changes', *Biology Direct*, 6: 26.

Clark, J. W., and Donoghue, P. C. J. 2017. 'Constraining the timing of whole genome duplication in plant evolutionary history', *Proceedings of the Royal Society B: Biological Sciences*, 284: 20170912.

Cohen, P. A., and Macdonald, F. A. 2015. 'The Proterozoic Record of Eukaryotes', *Paleobiology*, 41: 610-32.

Corliss, J. B. 1981. 'An hypothesis concerning the relationship between submarine hot springs and the origin of life on Earth', *Ocean. Acta*, 4: 59-69.

Cox, C. J., Foster, P. G., Hirt, R. P., Harris, S. R., and Embley, M. T. 2008. 'The archaebacterial origin of eukaryotes', *Proceedings of the National Academy of Sciences*, 105: 20356-61.

Criscuolo, A., and Gribaldo., S. 2011. 'Large-Scale Phylogenomic Analyses Indicate a Deep Origin of Primary Plastids within Cyanobacteria', *Molecular biology and evolution*, 28: 3019-32.

Crowe, S. A. et al. 2013. 'Atmospheric oxygenation three billion years ago', *Nature*, 501: 535.

Crowe, S. A. et al. 2008. 'Photoferrotrophs thrive in an Archean Ocean analogue', *Proceedings of the National Academy of Sciences*, 105: 15938-43.

Culver, S. J. 1991. 'Early Cambrian Foraminifera from West Africa', *Science*, 254: 689-91.

Cunningham, J. A., Liu, A. G., Bengtson, S., and Donoghue, P. C. J. 2017. 'The origin of animals: Can molecular clocks and the fossil record be reconciled?', *BioEssays*, 39: e201600120.

Cunningham, J. A. et al.. 2012. 'Distinguishing geology from biology in the Ediacaran Doushantuo biota relaxes constraints on the timing of the origin of bilaterians', *Proceedings of the Royal Society B: Biological Sciences*, 279: 2369-76.

Czaja, A. D. et al. 'Evidence for free oxygen in the Neoarchean ocean based on coupled iron–molybdenum isotope fractionation', *Geochimica et Cosmochimica Acta*, 86: 118-37.

Da Cunha, V., Gaia, M., Gadelle, D., Nasir, A., and Forterre, P. 2017. 'Lokiarchaea are close relatives of Euryarchaeota, not bridging the gap between prokaryotes and eukaryotes', *PLoS genetics*, 13: e1006810.

Davín, A. A. et al. 2018. 'Gene transfers can date the tree of life', *Nature Ecology & Evolution*, 2: 904-09.

Degnan, J. H., and Rosenberg, N. A. 2009. 'Gene tree discordance, phylogenetic inference and the multispecies coalescent', *Trends in Ecology & Evolution*, 24: 332-40.

Deusch, O. et al. 2008. 'Genes of Cyanobacterial Origin in Plant Nuclear Genomes Point to a Heterocyst-Forming Plastid Ancestor', *Molecular biology and evolution*, 25: 748-61.

Dodd, M. S. et al. 2017. 'Evidence for early life in Earth's oldest hydrothermal vent precipitates', *Nature*, 543: 60.

Dohrmann, M., and Wörheide, G. 2017. 'Dating early animal evolution using phylogenomic data', *Scientific reports*, 7: 3599.

dos Reis, M., Donoghue, P. C. J., and Yang, Z. 2015. 'Bayesian molecular clock dating of species divergences in the genomics era', *Nature Reviews Genetics*, 17: 71.

Drummond, A. J., Ho, S. Y. W., Phillips, M. J., and Rambaut, A. 2006. 'Relaxed Phylogenetics and Dating with Confidence', *PLOS Biology*, 4: e88.

Duda, J. et al. 2016. 'A Rare Glimpse of Paleoarchean Life: Geobiology of an Exceptionally Preserved Microbial Mat Facies from the 3.4 Ga Strelley Pool Formation, Western Australia', *PLOS ONE*, 11: e0147629.

Dunn, F. S., Liu, A. G., and Donoghue, P. C. J. 2018. 'Ediacaran developmental biology', *Biological Reviews*, 93: 914-32.

Dworkin, M., Falkow, S., Rosenberg, E., Schleifer, K. H., and Stackebrandt, E. 2006. 'The Prokaryotes.' in (Springer).

Edgar, R. C. 2004. 'MUSCLE: multiple sequence alignment with high accuracy and high throughput', *Nucleic Acids Research*, 32: 1792-97.

Embley, T. M., and Martin, W. 2006. 'Eukaryotic evolution, changes and challenges', *Nature*, 440: 623.

Eme, L., Sharpe, S. C., Brown, M. W, and Roger, A. J. 2014. 'On the age of eukaryotes: evaluating evidence from fossils and molecular clocks', *Cold Spring Harbor Perspectives in Biology*, 6: a016139.

Engel, A. E. J. et al.. 1968. 'Alga-Like Forms in Onverwacht Series, South Africa: Oldest Recognized Lifelike Forms on Earth', *Science*, 161: 1005-08.

Erwin, D. H. et al. 2011. 'The Cambrian Conundrum: Early Divergence and Later Ecological Success in the Early History of Animals', *Science*, 334: 1091-97.

Esser, C. et al 2004. 'A Genome Phylogeny for Mitochondria Among α-Proteobacteria and a Predominantly Eubacterial Ancestry of Yeast Nuclear Genes', *Molecular biology and evolution*, 21: 1643-60.

Ettema, T. J. G., Lindås, A. and Bernander, R. 2011. 'An actin-based cytoskeleton in archaea', *Molecular Microbiology*, 80: 1052-61.

Evans, P. N. et al. 2015. 'Methane metabolism in the archaeal phylum Bathyarchaeota revealed by genome-centric metagenomics', *Science*, 350: 434-38.

Fani, R., Liò, P., Chiarelli, I., and Bazzicalupo, M. 1994. 'The evolution of the histidine biosynthetic genes in prokaryotes: A common ancestor for the hisA and hisF genes', *Journal of Molecular Evolution*, 38: 489-95.

Fedonkin, M. A., Gehling, J. G., Grey, K., Narbonne, G. M., and Vickers-Rich, P. 2007. *The rise of animals: evolution and diversification of the kingdom Animalia* (JHU Press).

Fitzpatrick, David A., Christopher J. Creevey, and James O. McInerney. 2005. 'Genome Phylogenies Indicate a Meaningful α-Proteobacterial Phylogeny and Support a Grouping of the Mitochondria with the Rickettsiales', *Molecular biology and evolution*, 23: 74-85.

Forterre, P., and Philippe, H. 1999. 'Where is the root of the universal tree of life?', *BioEssays*, 21: 871-79.

Furukawa, R., Nakagawa, M., Kuroyanagi, T., Yokobori, S., and Yamagish, A.. 2017. 'Quest for Ancestors of Eukaryal Cells Based on Phylogenetic Analyses of Aminoacyl-tRNA Synthetases', *Journal of Molecular Evolution*, 84: 51-66.

Gavryushkina, A., Welch, D., Stadler, T., and Drummond, A. J. 2014. 'Bayesian Inference of Sampled Ancestor Trees for Epidemiology and Fossil Calibration', *PLOS Computational Biology*, 10: e1003919.

Gibson, T. M. et al. 2017. 'Precise age of Bangiomorpha pubescens dates the origin of eukaryotic photosynthesis', *Geology*, 46: 135-38.

Gillespie, J. H. 1991. *The causes of molecular evolution* (Oxford University Press: New York).

Gogarten, J. P. et al. 1989. 'Evolution of the vacuolar H+-ATPase: implications for the origin of eukaryotes', *Proceedings of the National Academy of Sciences*, 86: 6661-65.

Golubic, S., and Hofmann, H. J. 1976. 'Comparison of Holocene and Mid-Precambrian Entophysalidaceae (Cyanophyta) in Stromatolitic Algal Mats: Cell Division and Degradation', *Journal of Paleontology*, 50: 1074-82.

Gradstein, F. M, Ogg, J. G., Schmitz, M., and Ogg, G. 2012. *The geologic time scale 2012* (elsevier).

Gray, M. W. 2012. 'Mitochondrial Evolution', *Cold Spring Harbor Perspectives in Biology*, 4.

Gray, M. W. 2015. 'Mosaic nature of the mitochondrial proteome: Implications for the origin and evolution of mitochondria', *Proceedings of the National Academy of Sciences*, 112: 10133-38.

Gribaldo, S., and Cammarano, P. 1998. 'The root of the universal tree of life inferred from anciently duplicated genes encoding components of the protein-targeting machinery', *Journal of Molecular Evolution*, 47: 508-16.

Gross, J., and Bhattacharya, D. 2010. 'Uniting sex and eukaryote origins in an emerging oxygenic world', *Biology Direct*, 5: 53.

Grotzinger, J. P., and Rothman, D. H. 1996. 'An abiotic model for stromatolite morphogenesis', *Nature*, 383: 423-25.

Guy, L., and Ettema, T. J. G. 2011. 'The archaeal 'TACK' superphylum and the origin of eukaryotes', *Trends in Microbiology*, 19: 580-87.

Haldane, J. B. S. 1929. 'The origin of life: Rationalist Annual, v. 148'.

Halliday, A. N. 2008. 'A young Moon-forming giant impact at 70-110 million years accompanied by late-stage mixing, core formation and degassing of the Earth', *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 366: 4163-81.

Halliday, A. N., Rehkämper, M., Lee, D., and Yi, W. 1996. 'Early evolution of the Earth and Moon: new constraints from Hf-W isotope geochemistry', *Earth and Planetary Science Letters*, 142: 75-89.

Halliday, A. N. 2014. 'The origin and earliest history of the Earth', *Planets, Asteriods, Comets and The Solar System*: 149-211.

Han, T. M., and Runnegar, B. 1992. 'Megascopic eukaryotic algae from the 2.1-billion-year-old negaunee iron-formation, Michigan', *Science*, 257: 232-35.

Hanan, B. B., and Tilton, G. R. 1987. '60025: relict of primitive lunar crust?', *Earth and Planetary Science Letters*, 84: 15-21.

Harrison, T. M., Bell, E. A., and Boehnke, P. 2017. 'Hadean Zircon Petrochronology', *Reviews in Mineralogy and Geochemistry*, 83: 329-63.

Hayes, J. M. 1994. "Early life on earth." In *Nobel symposium*, 220-36.

Heaman, L. M., Le Cheminant, A. N., and Rainbird, R. H. 1992. 'Nature and timing of Franklin igneous events, Canada: Implications for a Late Proterozoic mantle plume and the break-up of Laurentia', *Earth and Planetary Science Letters*, 109: 117-31.

Heath, T. A., Huelsenbeck, J. P., and Stadler, T. 2014. 'The fossilized birth–death process for coherent calibration of divergence-time estimates', *Proceedings of the National Academy of Sciences*, 111: E2957-E66.

Hedges, S. B., Blair, J. E., Venturi, M. L., and Shoe, J. L. 2004. 'A molecular timescale of eukaryote evolution and the rise of complex multicellular life', *BMC Evolutionary Biology*, 4: 2.

Hickman, A. H. 2008. 'Regional review of the 3426–3350 Ma Strelley Pool Formation, Pilbara Craton, Western Australia', *West Australia Geolog Surv Rec*, 2008: 15.

Hoek, C., Mann, D., Jahns, H. M., and Jahns, M. 1995. *Algae: an introduction to phycology* (Cambridge university press).

Hoekzema, R. S, Brasier, M. D., Dunn, F. S., and Liu, A. G. 2017. 'Quantitative study of developmental biology confirms Dickinsonia as a metazoan', *Proceedings of the Royal Society B: Biological Sciences*, 284: 20171348.

Hofmann, H. J., Grey, K., Hickman, A. H., and Thorpe, R. I. 1999. 'Origin of 3.45 Ga coniform stromatolites in Warrawoona Group, Western Australia', *GSA Bulletin*, 111: 1256-62.

Hofmann, H. J. 1976. 'Precambrian microflora, Belcher Islands, Canada: significance and systematics', *Journal of Paleontology*: 1040-73.

Holland, H. D. 2006. 'The oxygenation of the atmosphere and oceans', *Philosophical Transactions of the Royal Society B: Biological Sciences*, 361: 903-15.

Homann, M., Heubeck, C., Airo, A., and Tice, M. M. 2015. 'Morphological adaptations of 3.22 Ga-old tufted microbial mats to Archean coastal habitats (Moodies Group, Barberton Greenstone Belt, South Africa)', *Precambrian Research*, 266: 47-64.

Horita, J., and Berndt, M. E. 1999. 'Abiogenic methane formation and isotopic fractionation under hydrothermal conditions', *Science*, 285: 1055-57.

Horodyski, R. J. 1982. 'Problematic Bedding-Plane Markings from the Middle Proterozoic Appekunny Argillite, Belt Supergroup, Northwestern Montana', *Journal of Paleontology*, 56: 882-89.

Hug, L. A. et al. 2016. 'A new view of the tree of life', *Nature Microbiology*, 1: 16048.

Hurley, J. H., and Hanson., P. I. 2010. 'Membrane budding and scission by the ESCRT machinery: it's all in the neck', *Nature Reviews Molecular Cell Biology*, 11: 556.

Inoue, J., Donoghue, P. C. J., and Yang, Z. 2009. 'The Impact of the Representation of Fossil Calibrations on Bayesian Estimation of Species Divergence Times', *Systematic Biology*, 59: 74-89.

Ivarsson, M. et al. 2012. 'Fossilized fungi in subseafloor Eocene basalts', *Geology*, 40: 163-66.

Ivarsson, M., Bengtson, S., Skogby, H., Belivanova, V., and Marone, F. 2013. 'Fungal colonies in open fractures of subseafloor basalt', *Geo-Marine Letters*, 33: 233-43.

Iwabe, N., Kuma, K., Hasegawa, M., Osawa, S., and Miyata, T. 1989. 'Evolutionary relationship of archaebacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes', *Proceedings of the National Academy of Sciences*, 86: 9355-59.

Jacobson, S. A. et al. 2014. 'Highly siderophile elements in Earth's mantle as a clock for the Moon-forming impact', *Nature*, 508: 84.

Javaux, E. J., Knoll, A. H., and Walter, M. R. 2004. 'TEM evidence for eukaryotic diversity in mid-Proterozoic oceans', *Geobiology*, 2: 121-32.

Javaux, E. J. 2019. 'Challenges in evidencing the earliest traces of life', *Nature*, 572: 451-60.

Javaux, E. J., Knoll, A. H., and Walter, M. R. 2001. 'Morphological and ecological complexity in early eukaryotic ecosystems', *Nature*, 412: 66-69.

Javaux, E. J., and Leopt, K.. 2018. 'The Paleoproterozoic fossil record: Implications for the evolution of the biosphere during Earth's middle-age', *Earth-Science Reviews*, 176: 68-86.

Javaux, E. J., Marshall, C. P., and Bekke, A.. 2010. 'Organic-walled microfossils in 3.2-billion-year-old shallow-marine siliciclastic deposits', *Nature*, 463: 934.

Kah, L. C., Sherman, A. G., Narbonne, G. M., Knoll, A. H., and Kaufman, A. J. 1999. 'δ13C stratigraphy of the Proterozoic Bylot Supergroup, Baffin Island, Canada: implications for regional lithostratigraphic correlations', *Canadian Journal of Earth Sciences*, 36: 313-32.

Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A., and Jermiin, L. S. 2017. 'ModelFinder: fast model selection for accurate phylogenetic estimates', *Nature Methods*, 14: 587.

Kamo, S. L., and Davis, D. W. 1994. 'Reassessment of Archean crustal development in the Barberton Mountain Land, South Africa, based on U-Pb dating', *Tectonics*, 13: 167-92.

Karnkowska, A. et al. 2016. 'A Eukaryote without a Mitochondrial Organelle', *Current Biology*, 26: 1274-84.

Kendall, B. et al. 2010. 'Pervasive oxygenation along late Archaean ocean margins', *Nature Geoscience*, 3: 647.

Kishino, H., Thorne, J. L., and Bruno, W. J. 2001. 'Performance of a divergence time estimation method under a probabilistic model of rate evolution', *Molecular biology and evolution*, 18: 352-61.

Klein, C., Beukes, N. J., and Schopf. J. W. 1987. 'Filamentous microfossils in the early Proterozoic Transvaal Supergroup: their morphology, significance, and paleoenvironmental setting', *Precambrian Research*, 36: 81-94.

Kleine, T., Münker, C., Mezger, K., and Palme, H. 2002. 'Rapid accretion and early core formation on asteroids and the terrestrial planets from Hf–W chronometry', *Nature*, 418: 952-55.

Kleine, T., Palme, H., Mezger, K., and Halliday, A. N. 2005. 'Hf-W Chronometry of Lunar Metals and the Age and Early Differentiation of the Moon', *Science*, 310: 1671-74.

Knoll, A. H., Javaux, E. J., Hewitt, D., and Cohen, P. 2006. 'Eukaryotic organisms in Proterozoic oceans', *Philosophical Transactions of the Royal Society B: Biological Sciences*, 361: 1023-38.

Knoll, A. H., and Nowak, M. A. 2017. 'The timetable of evolution', *Science Advances*, 3: e1603076.

Knoll, A. H., Strother, P. K., and Rossi, S. 1988. 'Distribution and diagenesis of microfossils from the lower Proterozoic Duck Creek Dolomite, Western Australia', *Precambrian Research*, 38: 257-79.

Knoll, A. H. 2014. 'Paleobiological Perspectives on Early Eukaryotic Evolution', *Cold Spring Harbor Perspectives in Biology*, 6.

Koeberl, C. 2006. 'Impact Processes on the Early Earth', *Elements*, 2: 211-16.

Konhauser, K. O. et al. 2002. 'Could bacteria have formed the Precambrian banded iron formations?', *Geology*, 30: 1079-82.

Koonin, E. V. 2010. 'The origin and early evolution of eukaryotes in the light of phylogenomics', *Genome Biology*, 11: 209.

Koumandou, V. L. et al. 2013 ‘Molecular paleontology and complexity in the last eukaryotic common ancestor’ *Critical Reviews in Biochemistry and Molecular* Biolology, 48: 373–396

Ku, C. et al. 2015. 'Endosymbiotic origin and differential loss of eukaryotic genes', *Nature*, 524: 427.

Kück, P., and Meusemann, K. 2010. 'FASconCAT: Convenient handling of data matrices', *Molecular Phylogenetics and Evolution*, 56: 1115-18.

Kurland, C. G., Collins, L. J., and Penny, D. 2006. 'Genomics and the Irreducible Nature of Eukaryote Cells', *Science*, 312: 1011-14.

Labandeira, C. C. 2018. 'The Fossil History of Insect Diversity', *Insect Biodiversity: Science and Society*, 2: 723-88.

Labedan, B. et al. 1999. 'The evolutionary history of carbamoyltransferases: a complex set of paralogous genes was already present in the last universal common ancestor', *Journal of Molecular Evolution*, 49: 461-73.

Lake, J. A., Henderson, E., Oakes, M., and Clark, M. W. 1984. 'Eocytes: a new ribosome structure indicates a kingdom with a close relationship to eukaryotes', *Proceedings of the National Academy of Sciences*, 81: 3786-90.

Lamb, D. M., Awramik, S. M., Chapman, D. J., and Zhu, S. 2009. 'Evidence for eukaryotic diversification in the∼ 1800 million-year-old Changzhougou Formation, North China', *Precambrian Research*, 173: 93-104.

Lane, N., and Martin, W. 2010. 'The energetics of genome complexity', *Nature*, 467: 929.

Lanfear, R., Calcott, B., Ho, S. Y. W., and Guindon, S. 2012. 'PartitionFinder: Combined Selection of Partitioning Schemes and Substitution Models for Phylogenetic Analyses', *Molecular biology and evolution*, 29: 1695-701.

Lartillot, N., Lepage, T., and Blanquart, S. 2009. 'PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating', *Bioinformatics*, 25: 2286-88.

Lartillot, N., and Philippe, H. 2004. 'A Bayesian Mixture Model for Across-Site Heterogeneities in the Amino-Acid Replacement Process', *Molecular biology and evolution*, 21: 1095-109.

Lawson, F. S., Charlebois, R. L., and Dillon, J. A. 1996. 'Phylogenetic analysis of carbamoylphosphate synthetase genes: complex evolutionary history includes an internal

duplication within a gene which can root the tree of life', *Molecular biology and evolution*, 13: 970-77.

LeCheminant, A. N., and Heaman, L. M. 1989. 'Mackenzie igneous events, Canada: Middle Proterozoic hotspot magmatism associated with ocean opening', *Earth and Planetary Science Letters*, 96: 38-48.

Lee, M. S. Y., Soubrier, J., and Edgecombe, G. D., 2013. 'Rates of Phenotypic and Genomic Evolution during the Cambrian Explosion', *Current Biology*, 23: 1889-95.

Lee, R. E. 2008. *Phycology* (Cambridge University Press: New York).

Leiming, Y., Xunlai, Y., Fanwei, M., and Jie, H. 2005. 'Protists of the Upper Mesoproterozoic Ruyang Group in Shanxi Province, China', *Precambrian Research*, 141: 49-66.

Lenton, T. M., Boyle, R. A., Poulton, S. W., Shields-Zhou, G. A., and Butterfield, N. J., 2014. 'Co-evolution of eukaryotes and ocean oxygenation in the Neoproterozoic era', *Nature Geoscience*, 7: 257.

Lepage, T., Bryant, D., Philippe, H., and Lartillot, N. 2007. 'A General Comparison of Relaxed Molecular Clock Models', *Molecular biology and evolution*, 24: 2669-80.

Lepage, T., Lawi, S., Tupper, P., and Bryant, D. 2006. 'Continuous and tractable models for the variation of evolutionary rates', *Mathematical biosciences*, 199: 216-33.

Lepland, A., Arrhenius, G., and Cornell, D. 2002. 'Apatite in early Archean Isua supracrustal rocks, southern West Greenland: its origin, association with graphite and potential as a biomarker', *Precambrian Research*, 118: 221-41.

Lepot, K. et al. 2013. 'Texture-specific isotopic compositions in 3.4Gyr old organic matter support selective preservation in cell-like structures', *Geochimica et Cosmochimica Acta*, 112: 66-86.

Li, H. et al. 2013. 'Recent advances in the study of the Mesoproterozoic geochronology in the North China Craton', *Journal of Asian Earth Sciences*, 72: 216-27.

Linder, M., Britton, T., and Sennblad, B. 2011. 'Evaluation of Bayesian Models of Substitution Rate Evolution—Parental Guidance versus Mutual Independence', *Systematic Biology*, 60: 329-42.

Lipps, J. H. 1992. 'Proterozoic and Cambrian skeletonized protists', *The Proterozoic Biosphere*: 237-40.

Lipps, J. H., and Rozanov, A. Y. 1996. 'The late Precambrian-Cambrian agglutinated fossil Platysolenites', *Paleontological Journal*, 30: 679-87.

Loader, S. P. et al. 2007. 'Relative time scales reveal multiple origins of parallel disjunct distributions of African caecilian amphibians', *Biology Letters*, 3: 505-08.

Lollar, B. S., Westgate, T. D., Ward, J. A., Slater, G. F., and Lacrampe-Couloume, G. 2002. 'Abiogenic formation of alkanes in the Earth's crust as a minor source for global hydrocarbon reservoirs', *Nature*, 416: 522.

Long, D. G. F., and Turner, E. C. 2012. 'Tectonic, sedimentary and metallogenic re-evaluation of basal strata in the Mesoproterozoic Bylot basins, Nunavut, Canada: Are unconformity-type uranium concentrations a realistic expectation?', *Precambrian Research*, 214-215: 192-209.

Lopez, P., Forterre, P., and Philippe, H. 1999. 'The root of the tree of life in the light of the covarion model', *Journal of Molecular Evolution*, 49: 496-508.

Loron, C. C. et al. 2019. 'Early fungi from the Proterozoic era in Arctic Canada', *Nature*, 570: 232-35.

Lowe, D. R. 1994. 'Abiological origin of described stromatolites older than 3.2 Ga', *Geology*, 22: 387-90.

Lozano-Fernandez, J., dos Reis, M., Donoghue, P. C. J., and Pisani, D. 2017. 'RelTime Rates Collapse to a Strict Clock When Estimating the Timeline of Animal Diversification', *Genome biology and evolution*, 9: 1320-28.

Lu, S., Yang, C., and Zhu, S. 1996. *The Precambrian Continental Crust fromEastern Hebei to Jixian, Tianjin 30th International Geological Congress* (Geological Publishing House: Beijing).

Lyons, T. W., Reinhard, C. T., and Planavsky, N. J. 2014. 'The rise of oxygen in Earth’s early ocean and atmosphere', *Nature*, 506: 307.

Mark, D. F. et al. 2011. '40Ar/39Ar dating of hydrothermal activity, biota and gold mineralization in the Rhynie hot-spring system, Aberdeenshire, Scotland', *Geochimica et Cosmochimica Acta*, 75: 555-69.

Martijn, J., Vosseberg, J., Guy, L., Offre, P., and Ettema, T. J. G. 2018. 'Deep mitochondrial origin outside the sampled alphaproteobacteria', *Nature*, 557: 101-05.

Martin, W. F. et al. 2017. 'Late mitochondrial origin is an artifact', *Genome biology and evolution*, 9: 373-79.

Martin, W., and Russell, M. J. 2006. 'On the origin of biochemistry at an alkaline hydrothermal vent', *Philosophical Transactions of the Royal Society B: Biological Sciences*, 362: 1887-926.

Mayr, U. 2004. *Geology of Eastern Prince of Wales Island and Adjacent Smaller Islands, Nunavut (parts of NTS 68D, Baring Channel and 68A, Fisher Lake)* (Geological Survey of Canada).

McIlroy, D., Green, O. R., and Brasier, M. D. 2001. 'Palaeobiology and evolution of the earliest agglutinated Foraminifera: Platysolenites, Spirosolenites and related forms', *Lethaia*, 34: 13-29.

McInerney, J. O., O'Connell, M. J., and Pisani, D. 2014. 'The hybrid nature of the Eukaryota and a consilient view of life on Earth', *Nature Reviews Microbiology*, 12: 449.

McInerney, J., Pisani, D., and O'Connell, M. J. 2015. 'The ring of life hypothesis for eukaryote origins is supported by multiple kinds of data', *Philosophical Transactions of the Royal Society B: Biological Sciences*, 370: 20140323.

McLoughlin, N., Wilson, N. A., and Brasier, M. D. 2008. 'Growth of synthetic stromatolites and wrinkle structures in the absence of microbes–implications for the early fossil record', *Geobiology*, 6: 95-105.

Moczydłowska, M., Landing, E., Zang, W., and Palacios, T. 2011. 'Proterozoic phytoplankton and timing of Chlorophyte algae origins', *Palaeontology*, 54: 721-33.

Mojzsis, S. J. et al. 1996. 'Evidence for life on Earth before 3,800 million years ago', *Nature*, 384: 55.

Morris, J. L. et al. 2018. 'The timescale of early land plant evolution', *Proceedings of the National Academy of Sciences*, 115: E2274-E83.

Mulkidjanian, A. Y., Makarova, K. S., Galperin, M. Y., and Koonin, E. V. 2007. 'Inventing the dynamo machine: the evolution of the F-type and V-type ATPases', *Nature Reviews Microbiology*, 5: 892.

Narbonne, G. M., Xiao, S., Shields, G. A., and Gehling, J. G. 2012. 'The Ediacaran Period', *The geologic time scale*, 1: 413-35.

Nelson, D R. 2005. "178042: altered volcaniclastic sandstone, Table Top Well; Geochronology dataset 564." In *Compilation of geochronology data, June 2007 update*. Geological Survey of Western Australia.

Nguyen, L-T., Schmidt, H. A., von Haeseler, A., and Minh, B. Q. 2014. 'IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies', *Molecular biology and evolution*, 32: 268-74.

Nutman, A. P., Bennett, V. C., Friend, C. R. L., Van Kranendonk, M. J., and Chivas, A. R. 2016. 'Rapid emergence of life shown by discovery of 3,700-million-year-old microbial structures', *Nature*, 537: 535.

Nyunoya, H., and Lusty, C. J. 1983. 'The carB gene of Escherichia coli: a duplicated gene coding for the large subunit of carbamoyl-phosphate synthetase', *Proceedings of the National Academy of Sciences*, 80: 4629-33.

O'Reilly, J. E., and Donoghue, P. C. J. 2019. 'The Effect of Fossil Sampling on the Estimation of Divergence Times with the Fossilized Birth–Death Process', *Systematic Biology*.

Ochoa de Alda, J. A. G., Esteban, R., Diago, M. L., and Houmard, J. 2014. 'The plastid ancestor originated among one of the major cyanobacterial lineages', *Nature Communications*, 5: 4937.

Olson, S. L., Kump, L. R., and Kasting, J. F. 2013. 'Quantifying the areal extent and dissolved oxygen concentrations of Archean oxygen oases', *Chemical Geology*, 362: 35-43.

Pace, N. R., Olsen, G. J., and Woese, C. R. 1986. 'Ribosomal RNA phylogeny and the primary lines of evolutionary descent', *Cell*, 45: 325-26.

Parfrey, L. W., Lahr, D. J. G., Knoll, A. H., and Katz, L. A. 2011. 'Estimating the timing of early eukaryotic diversification with multigene molecular clocks', *Proceedings of the National Academy of Sciences*, 108: 13624-29.

Parham, J. F. et al. 2011. 'Best Practices for Justifying Fossil Calibrations', *Systematic Biology*, 61: 346-59.

Parry, S. F., Noble, S. R., Crowley, Q. G., and Wellman, C. H. 2013. 'Reply to Discussion on 'A high-precision U–Pb age constraint on the Rhynie Chert Konservat-Lagerstätte: time scale and other implications'', *Journal of the Geological Society,* 168: 863–872; 170: 703-06.

Parry, S. F., Noble, S. R., Crowley, Q. G., and Wellman, C. H. 2011. 'A high-precision U–Pb age constraint on the Rhynie Chert Konservat-Lagerstätte: time scale and other implications', *Journal of the Geological Society*, 168: 863-72.

Peng, Y., Bao, H., and Yuan, X. 2009. 'New morphological observations for Paleoproterozoic acritarchs from the Chuanlinggou Formation, North China', *Precambrian Research*, 168: 223-32.

Peters, S. E., and McClennen, M. 2016. 'The Paleobiology Database application programming interface', *Paleobiology*, 42: 1-7.

Peterson, K. J., and Butterfield, N. J. 2005. 'Origin of the Eumetazoa: Testing ecological predictions of molecular clocks against the Proterozoic fossil record', *Proceedings of the National Academy of Sciences of the United States of America*, 102: 9547-52.

Pflug, H. D., and Jaeschke-Boyer, H. 1979. 'Combined structural and chemical analysis of 3,800-Myr-old microfossils', *Nature*, 280: 483.

Philippe, H., and Forterre, P. 1999. 'The rooting of the universal tree of life is not reliable', *Journal of Molecular Evolution*, 49: 509-23.

Pisani, D., Cotton, J. A., and McInerney, J. O. 2007. 'Supertrees Disentangle the Chimerical Origin of Eukaryotic Genomes', *Molecular biology and evolution*, 24: 1752-60.

Pisani, D., and Liu, A. G. 2015. 'Animal Evolution: Only Rocks Can Set the Clock', *Current Biology*, 25: R1079-R81.

Pittis, A. A, and Gabaldón, T. 2016. 'Late acquisition of mitochondria by a host with chimaeric prokaryotic ancestry', *Nature*, 531: 101.

Planavsky, N. J. et al. 2014. 'Evidence for oxygenic photosynthesis half a billion years before the Great Oxidation Event', *Nature Geoscience*, 7: 283.

Ponce-Toledo, R. I. et al. 2017. 'An Early-Branching Freshwater Cyanobacterium at the Origin of Plastids', *Current Biology*, 27: 386-91.

Porter, S. M., Meisterfeld, R., and Knoll, A. H. 2003. 'Vase-shaped microfossils from the Neoproterozoic Chuar Group, Grand Canyon: a classification guided by modern testate amoebae', *Journal of Paleontology*, 77: 409-29.

Rambaut, A., Drummond, A. J., Xie, D., Baele, G., and Suchard, M. A. 2018. 'Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7', *Systematic Biology*, 67: 901-04.

Rannala, B., and Yang, Z. 2007. 'Inferring speciation times under an episodic molecular clock', *Systematic Biology*, 56: 453-66.

Redecker, D., Kodner, R. and Graham, L. E. 2000. 'Glomalean Fungi from the Ordovician', *Science*, 289: 1920-21.

Reisz, R. R., and Muller, J. 2005. 'Molecular timescales and the fossil_record: a paleontological perspective.' *Trends in Genetics.* 20: 237–241

Rice, C. M., and Ashcroft, W. A. 2003. 'The geology of the northern half of the Rhynie Basin, Aberdeenshire, Scotland', *Transactions of the Royal Society of Edinburgh: Earth Sciences*, 94: 299-308.

Riding, R., Fralick, P., and Liang, L. 2014. 'Identification of an Archean marine oxygen oasis', *Precambrian Research*, 251: 232-37.

Rivera, M. C., and Lake, J. A. 2004. 'The ring of life provides evidence for a genome fusion origin of eukaryotes', *Nature*, 431: 152-55.

Rivera, M. C., and Lake, J. A. 1992. 'Evidence that eukaryotes and eocyte prokaryotes are immediate relatives', *Science*, 257: 74-76.

Rodríguez-Ezpeleta, N., and Embley, T. M. 2012. 'The SAR11 Group of Alpha-Proteobacteria Is Not Related to the Origin of Mitochondria', *PLOS ONE*, 7: e30520.

Roger, A. J., Muñoz-Gómez, S. A., and Kamikawa, R. 2017. 'The Origin and Diversification of Mitochondria', *Current Biology*, 27: R1177-R92.

Ronquist, F. et al. 2012. 'A total-evidence approach to dating with fossils, applied to the early radiation of the Hymenoptera', *Systematic Biology*, 61: 973-99.

Ronquist, F. et al. 2012. 'MrBayes 3.2: Efficient Bayesian Phylogenetic Inference and Model Choice Across a Large Model Space', *Systematic Biology*, 61: 539-42.

Rosing, M. T. 1999. '13C-depleted carbon microparticles in> 3700-Ma sea-floor sedimentary rocks from West Greenland', *Science*, 283: 674-76.

Rozanov, A. Y. 1983. 'Platysolenites', *Upper Precambrian and Cambrian palaeontology of the East European Platform: Warsaw, Wydawnictwa Geologiczne*: 94-100.

Rozanov, A. Y., Zhuravlev, A. Y., Lipps, J. H., and Signor, P. W. 1992. "Origin and Early Evolution of the Metazoa." In.: Plenum Press New York.

Runnegar, B., Dollase, W. A., Ketcham, R. A., Colbert, M., and Carlson, W. D. 2001. "Early Archean sulfates from Western Australia first formed as hydrothermal barites not gypsum evaporites." In *Geol. Soc. Am. Abstracts with Programs*.

Russell, M. J., Hall, A. J., and Martin, W. F. 2010. 'Serpentinization as a source of energy at the origin of life', *Geobiology*, 8: 355-71.

Ryder, G. 2002. 'Mass flux in the ancient Earth-Moon system and benign implications for the origin of life on Earth', *Journal of Geophysical Research: Planets*, 107: 6-1-6-13.

Sánchez-Baracaldo, P., Raven, J. R., Pisani, D., and Knoll, A. H. 2017. 'Early photosynthetic eukaryotes inhabited low-salinity habitats', *Proceedings of the National Academy of Sciences*, 114: E7737-E45.

Sanderson, M. J. 1997. 'A nonparametric approach to estimating divergence times in the absence of rate constancy'.

Satkoski, A. M, Beukes, N. J., Li, W., Beard, B. L., and Johnson, C. L. 2015. 'A redox-stratified ocean 3.2 billion years ago', *Earth and Planetary Science Letters*, 430: 43-53.

Schidlowski, M. 1988. 'A 3,800-million-year isotopic record of life from carbon in sedimentary rocks', *Nature*, 333: 313.

Schirrmeister, B. E., de Vos, J. M., Antonelli, A., and Bagheri, H. C. 2013. 'Evolution of multicellularity coincided with increased diversification of cyanobacteria and the Great Oxidation Event', *Proceedings of the National Academy of Sciences*, 110: 1791-96.

Schirrmeister, B. E., Gugger, M., and Donoghue, P. C. J. 2015. 'Cyanobacteria and the Great Oxidation Event: evidence from genes and fossils', *Palaeontology*, 58: 769-85.

Schopf, J. W. 1993. 'Microfossils of the Early Archean Apex Chert: New Evidence of the Antiquity of Life', *Science*, 260: 640-46.

Schopf, J. W. 2006. 'Fossil evidence of Archaean life', *Philosophical Transactions of the Royal Society B: Biological Sciences*, 361: 869-85.

Schumann, G., Manz, W., Reitner, J., and Lustrino, M. 2004. 'Ancient Fungal Life in North Pacific Eocene Oceanic Crust', *Geomicrobiology Journal*, 21: 241-46.

Schwartz, R. M., and Dayhoff, M. O. 1978. 'Origins of Prokaryotes, Eukaryotes, Mitochondria, and Chloroplasts', *Science*, 199: 395-403.

Sharma, M., and Shukla, Y. 2009 ‘Taxonomy and affinity of Early Mesoproterozoic megascopic helically coiled and related fossils from the Rohtas Formation, the Vindhyan Supergroup, India.’ *Precambrian Research*, 173.1-4: 105-122.

Shen, Y., Buick, R., and Canfield, D. E. 2001. 'Isotopic evidence for microbial sulphate reduction in the early Archaean era', *Nature*, 410: 77-81.

Shih, P. M, and Matzke, N. J. 2013. 'Primary endosymbiosis events date to the later Proterozoic with cross-calibrated phylogenetic dating of duplicated ATPase proteins', *Proceedings of the National Academy of Sciences*, 110: 12355-60.

Solow, A. R., and Smith, W. 1997. 'On fossil preservation and the stratigraphic ranges of taxa', *Paleobiology*, 23: 271-77.

Sousa, F. L., Nelson-Sathi, S., and Martin, W. F. 2016. 'One step beyond a ribosome: The ancient anaerobic core', *Biochimica et Biophysica Acta (BBA) - Bioenergetics*, 1857: 1027-38.

Spang, A. et al. 2018. 'Asgard archaea are the closest prokaryotic relatives of eukaryotes', *PLoS genetics*, 14: e1007080.

Spang, A. et al. 2015. 'Complex archaea that bridge the gap between prokaryotes and eukaryotes', *Nature*, 521: 173.

Stadler, T. 2010. 'Sampling-through-time in birth–death trees', *Journal of theoretical biology*, 267: 396-404.

Stadler, T., Kühnert, D., Bonhoeffer, S., and Drummond, A. J. 2013. 'Birth–death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis C virus (HCV)', *Proceedings of the National Academy of Sciences*, 110: 228-33.

Sugitani, K., Mimura, K., Takeuchi, M., Lepot, K., Ito, S., and Javaux, E. J. 2015. 'Early evolution of large micro-organisms with cytological complexity revealed by microanalyses of 3.4 Ga organic-walled microfossils', *Geobiology*, 13: 507-21.

Sugitani, K. et al. 2015. 'A Paleoarchean coastal hydrothermal field inhabited by diverse microbial communities: the Strelley Pool Formation, Pilbara Craton, Western Australia', *Geobiology*, 13: 522-45.

Sugitani, K. et al. 2010. 'Biogenicity of Morphologically Diverse Carbonaceous Microstructures from the ca. 3400 Ma Strelley Pool Formation, in the Pilbara Craton, Western Australia', *Astrobiology*, 10: 899-920.

Sugitani, K., Mimura, K., Nagaoka, T., Lepot, K., and Takeuchi, M. 2013. 'Microfossil assemblage from the 3400 Ma Strelley Pool Formation in the Pilbara Craton, Western Australia: results form a new locality', *Precambrian Research*, 226: 59-74.

Szathmáry, E., and Smith, J. M. 1995. 'The major evolutionary transitions', *Nature*, 374: 227-32.

Taylor, T. N., Hass, H., and Kerp, H. 1999. 'The oldest fossil ascomycetes', *Nature*, 399: 648-48.

Taylor, T. N., Hass, H., and Kerp, H., Krings, M., and Hanlin, R. T. 2005. 'Perithecial ascomycetes from the 400 million year old Rhynie chert: an example of ancestral polymorphism', *Mycologia*, 97: 269-85.

Team, R Core. 2013. 'R: A language and environment for statistical computing'.

Tera, F., Papanastassiou, D. A., and Wasserburg, G. J. 1974. "The lunar time scale and a summary of isotopic evidence for a terminal lunar cataclysm." In *Lunar and Planetary Science Conference*.

Thiergart, T., Landan, G., Schenk, M., Dagan, T., and Martin, W. F. 2012. 'An Evolutionary Network of Genes Present in the Eukaryote Common Ancestor Polls Genomes on Eukaryotic and Mitochondrial Origin', *Genome biology and evolution*, 4: 466-85.

Thorne, J. L., Kishino, H., and Painter, I. S.. 1998. 'Estimating the rate of evolution of the rate of molecular evolution', *Molecular biology and evolution*, 15: 1647-57.

Thorne, J. L., Kishino, H. 2002. 'Divergence time and evolutionary rate estimation with multilocus data', *Systematic Biology*, 51: 689-702.

Tomitani, A., Knoll, A. H., Cavanaugh, C. M., and Ohno, T. 2006. 'The evolutionary diversification of cyanobacteria: molecular-phylogenetic and paleontological perspectives', *Proceedings of the National Academy of Sciences of the United States of America*, 103: 5442-47.

Touboul, M., Kleine, T., Bourdon, B., Palme, H., and Wieler, R. 2007. 'Late formation and prolonged differentiation of the Moon inferred from W isotopes in lunar metals', *Nature*, 450: 1206.

Tria, F. D. K., Landan, G., and Dagan, T. 2017. 'Phylogenetic rooting using minimal ancestor deviation', *Nature Ecology &Amp; Evolution*, 1: 0193.

Turner, E. C., and Kamber, B. S. 2012. 'Arctic Bay Formation, Borden Basin, Nunavut (Canada): Basin evolution, black shale, and dissolved metal systematics in the Mesoproterozoic ocean', *Precambrian Research*, 208-211: 1-18.

Ueno, Y. 2001. 'Early Archean (ca. 3.5 Ga) microfossils and ^{13}C-depleted carbonaceous matter in the North Pole area, Western Australia : Field occurrence and geochemistry', *Geochemistry and the origin of life*: 203-36.

Ueno, Y., Yamada, K., Yoshida, N., Maruyama, S., and Isozaki, Y. 2006. 'Evidence from fluid inclusions for microbial methanogenesis in the early Archaean era', *Nature*, 440: 516-19.

Valley, J. W., Peck, W. H., King, E. M., and Wilde, S. A. 2002. 'A cool early Earth', *Geology*, 30: 351-54.

Van Kranendonk, M. J. et al. 2012. 'A Chronostratigraphic Division of the Precambrian: Possibilities and Challenges.' in, *The Geologic Time Scale 2012* (Elsevier BV).

Van Kranendonk, M. J. 2006. 'Volcanic degassing, hydrothermal circulation and the flourishing of early life on Earth: A review of the evidence from c. 3490-3240 Ma rocks of the Pilbara Supergroup, Pilbara Craton, Western Australia', *Earth-Science Reviews*, 74: 197-240.

Van Zuilen, M. A., Lepland, A., and Arrhenius, G. 2002. 'Reassessing the evidence for the earliest traces of life', *Nature*, 418: 627.

Van Zuilen, Mark A. et al. 2003. 'Graphite and carbonates in the 3.8 Ga old Isua Supracrustal Belt, southern West Greenland', *Precambrian Research*, 126: 331-48.

Vanwonterghem, I. et al. 2016. 'Methylotrophic methanogenesis discovered in the archaeal phylum Verstraetearchaeota', *Nature Microbiology*, 1: 16170.

Villeneuve, M., Theveniaut, H., Ndiaye, P. M., and Retière, S. 2014. 'Re-assessment of the northern Guinean “Koubia–Lessere unconformity” (KLU): Consequences on the geological correlations throughout West Africa', *Comptes Rendus Geoscience*, 346: 262-72.

Wacey, D. 2009. *Early life on earth: a practical guide* (Springer Science & Business Media).

Wacey, D. 2010. 'Stromatolites in the~ 3400 Ma Strelley Pool Formation, Western Australia: examining biogenicity from the macro-to the nano-scale', *Astrobiology*, 10: 381-95.

Wacey, D., Kilburn, M. R., Saunders, M., Cliff, J., and Brasier, M. D. 2011. 'Microfossils of sulphur-metabolizing cells in 3.4-billion-year-old rocks of Western Australia', *Nature Geoscience*, 4: 698.

Wacey, D., McLoughlin, N., Whitehouse, M. J., and Kilburn, M. R. 2010. 'Two coexisting sulfur metabolisms in a ca. 3400 Ma sandstone', *Geology*, 38: 1115-18.

Walsh, M. M., and Lowe, D. R. 1985. 'Filamentous microfossils from the 3,500-Myr-old Onverwacht Group, Barberton Mountain Land, South Africa', *Nature*, 314: 530-32.

Walter, M. R., Buick, R., and Dunlop, J. S. R. 1980. 'Stromatolites 3,400–3,500 Myr old from the North Pole area, Western Australia', *Nature*, 284: 443-45.

Wang, Z., and Wu, M. 2014. 'Phylogenomic Reconstruction Indicates Mitochondrial Ancestor Was an Energy Parasite', *PLOS ONE*, 9: e110685.

Wang, Z., and Wu, M. 2015. 'An integrated phylogenomic approach toward pinpointing the origin of mitochondria', *Scientific reports*, 5: 7949.

Warnock, R. C. M., Parham, J. F., Joyce, W. G., Lyson, T. R., and Donoghue, P. C. J. 2015. 'Calibration uncertainty in molecular dating analyses: there is no substitute for the prior evaluation of time priors', *Proceedings of the Royal Society B: Biological Sciences*, 282: 20141013.

Warnock, R. C. M., Yang, Z., and Donoghue, P. C. J. 2012. 'Exploring uncertainty in the calibration of the molecular clock', *Biology Letters*, 8: 156-59.

Weiss, M. C., Preiner, M., Xavier, J. C., Zimorski, V., and Martin, W. F. 2018. 'The last universal common ancestor between ancient Earth chemistry and the onset of genetics', *PLoS genetics*, 14: e1007518.

Weiss, M. C. et al.. 2016. 'The physiology and habitat of the last universal common ancestor', *Nature Microbiology*, 1: 16116.

Wellman, C. H. 2006. 'Spore assemblages from the Lower Devonian 'Lower Old Red Sandstone' deposits of the Rhynie outlier, Scotland', *Transactions of the Royal Society of Edinburgh: Earth Sciences*, 97: 167-211.

Westall, F. et al. 2001. 'Early Archean fossil bacteria and biofilms in hydrothermally-influenced sediments from the Barberton greenstone belt, South Africa', *Precambrian Research*, 106: 93-116.

Westall, F. et al. 2006. 'Implications of a 3.472-3.333 Gyr-old subaerial microbial mat from the Barberton greenstone belt, South Africa for the UV environmental conditions on the early Earth', *Philosophical Transactions of the Royal Society B: Biological Sciences*, 361: 1857-76.

Williams, K. P., Sobral, B. W., and Dickerman, A. W. 2007. 'A robust species tree for the alphaproteobacteria', *Journal of bacteriology*, 189: 4578-86.

Williams, T. A., Foster, P. G., Cox, C. J., and Embley, T. M. 2013. 'An archaeal origin of eukaryotes supports only two primary domains of life', *Nature*, 504: 231.

Williams, T. A., Foster, P. G., Nye, T. M. W, Cox, C. J., and Embley, T. M. 2012. 'A congruent phylogenomic signal places eukaryotes within the Archaea', *Proceedings of the Royal Society B: Biological Sciences*, 279: 4870-79.

Williams, T. A. et al. 2017. 'Integrative modeling of gene and genome evolution roots the archaeal tree of life', *Proceedings of the National Academy of Sciences*, 114: E4602-E11.

Woese, C. R., Kandler, O., and Wheelis, M. L. 1990. 'Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya', *Proceedings of the National Academy of Sciences*, 87: 4576-79.

Woese, C. R., and Fox, G. E. 1977. 'Phylogenetic structure of the prokaryotic domain: The primary kingdoms', *Proceedings of the National Academy of Sciences*, 74: 5088-90.

Wolfe, J. M., and Fournier, G. P. 2018. 'Horizontal gene transfer constrains the timing of methanogen evolution', *Nature Ecology & Evolution*, 2: 897-903.

Xiao, S., Knoll, A. H., Yuan, X., and Pueschel, C. M. 2004. 'Phosphatized multicellular algae in the Neoproterozoic Doushantuo Formation, China, and the early evolution of florideophyte red algae', *American Journal of Botany*, 91: 214-27.

Yang, D., Oyaizu, Y., Oyaizu, H., Olsen, G. J., and Woese, C. R. 1985. 'Mitochondrial origins', *Proceedings of the National Academy of Sciences of the United States of America*, 82: 4443-47.

Yang, E. C. 2016. 'Divergence time estimates and the evolution of major lineages in the florideophyte red algae', *Scientific reports*, 6: 21361.

Yang, Z. 2007. 'PAML 4: Phylogenetic Analysis by Maximum Likelihood', *Molecular biology and evolution*, 24: 1586-91.

Yang, Z., and Rannala, B. 2006. 'Bayesian estimation of species divergence times under a molecular clock using multiple fossil calibrations with soft bounds', *Molecular biology and evolution*, 23: 212-26.

Yin, L. 1997. 'Acanthomorphic acritarchs from Meso-Neoproterozoic shales of the Ruyang Group, Shanxi, China', *Review of Palaeobotany and Palynology*, 98: 15-25.

Yin, L., and Yuan, X. 2003. 'Review of the microfossil assemblage rom the late mesoproterozoic ruyang group in Shanxi, China', *Acta Micropalaeontologica Sinica*, 21: 39-46.

Yuan, X., Xiao, S., and Taylor, T. N. 2005. 'Lichen-Like Symbiosis 600 Million Years Ago', *Science*, 308: 1017-20.

Zaremba-Niedzwiedzka, K. et al. 2017. 'Asgard archaea illuminate the origin of eukaryotic cellular complexity', *Nature*, 541: 353.

Zhaxybayeva, O., Lapierre, P., and Gogarten, J. P. 2005. 'Ancient gene duplications and the root (s) of the tree of life', *Protoplasma*, 227: 53-64.

Zhongying, Z. 1986. 'Clastic facies microfossils from the Chuanlinggou Formation (1800 Ma) near Jixian, North China', *Journal of Micropalaeontology*, 5: 9-16.

Zhu, S. et al. 2016. 'Decimetre-scale multicellular eukaryotes from the 1.56-billion-year-old Gaoyuzhuang Formation in North China', *Nature Communications*, 7: 11500.

Zuckerkandl, E., and Pauling, L. 1962. *Molecular disease, evolution, and genetic heterogeneity.* (Academic Press.: New York).

Zuckerkandl, E., and Pauling, L. 1965. 'Evolutionary divergence and convergence in proteins.' in, *Evolving genes and proteins* (Elsevier).

Zúñiga, M., Pérez,G., and González-Candelas, F. 2002. 'Evolution of arginine deiminase (ADI) pathway genes', *Molecular Phylogenetics and Evolution*, 25: 429-44.

Appendix A

Supplementary Figures

List of Appendix Figures

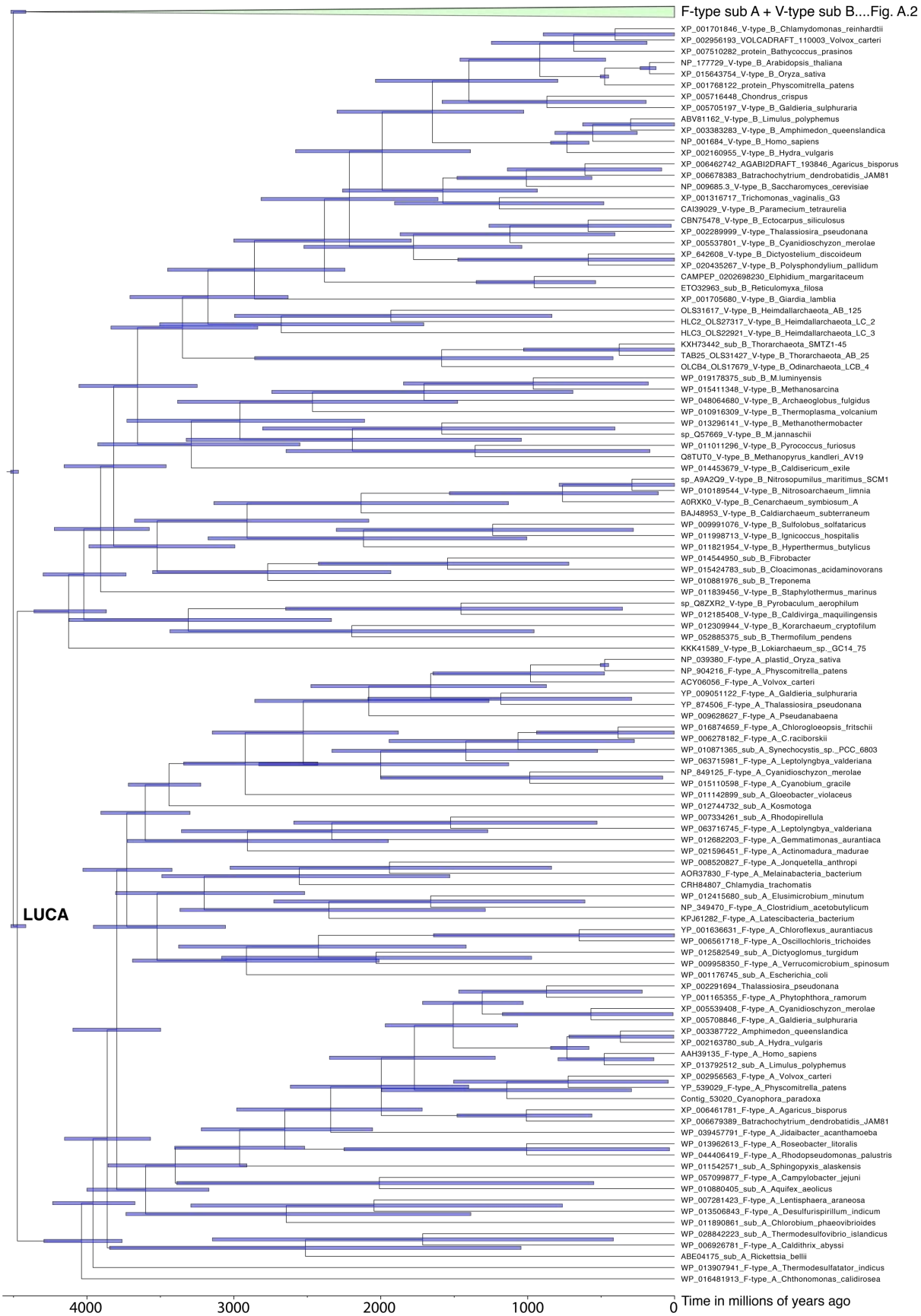


Figure A.1. Divergence time tree of the F-type and V-type ATPase gene family with a focus on F-type sub A and V-type sub B genes. Blue bars indicate the 95% credible intervals.

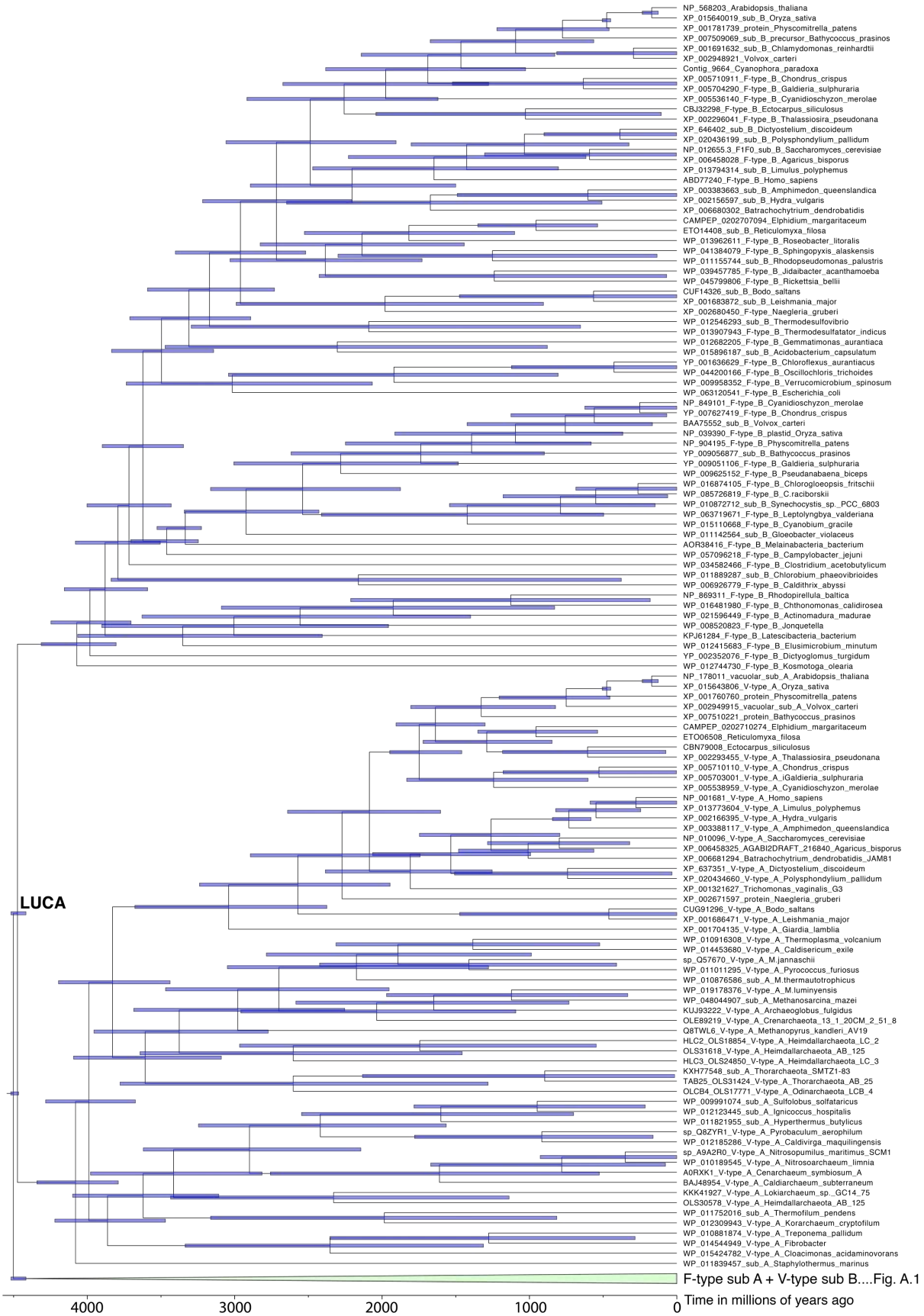


Figure A.2. Divergence time tree of the F-type and V-type ATPase gene family with a focus on F-type sub B and V-type sub A genes. Blue bars indicate the 95% credible intervals.

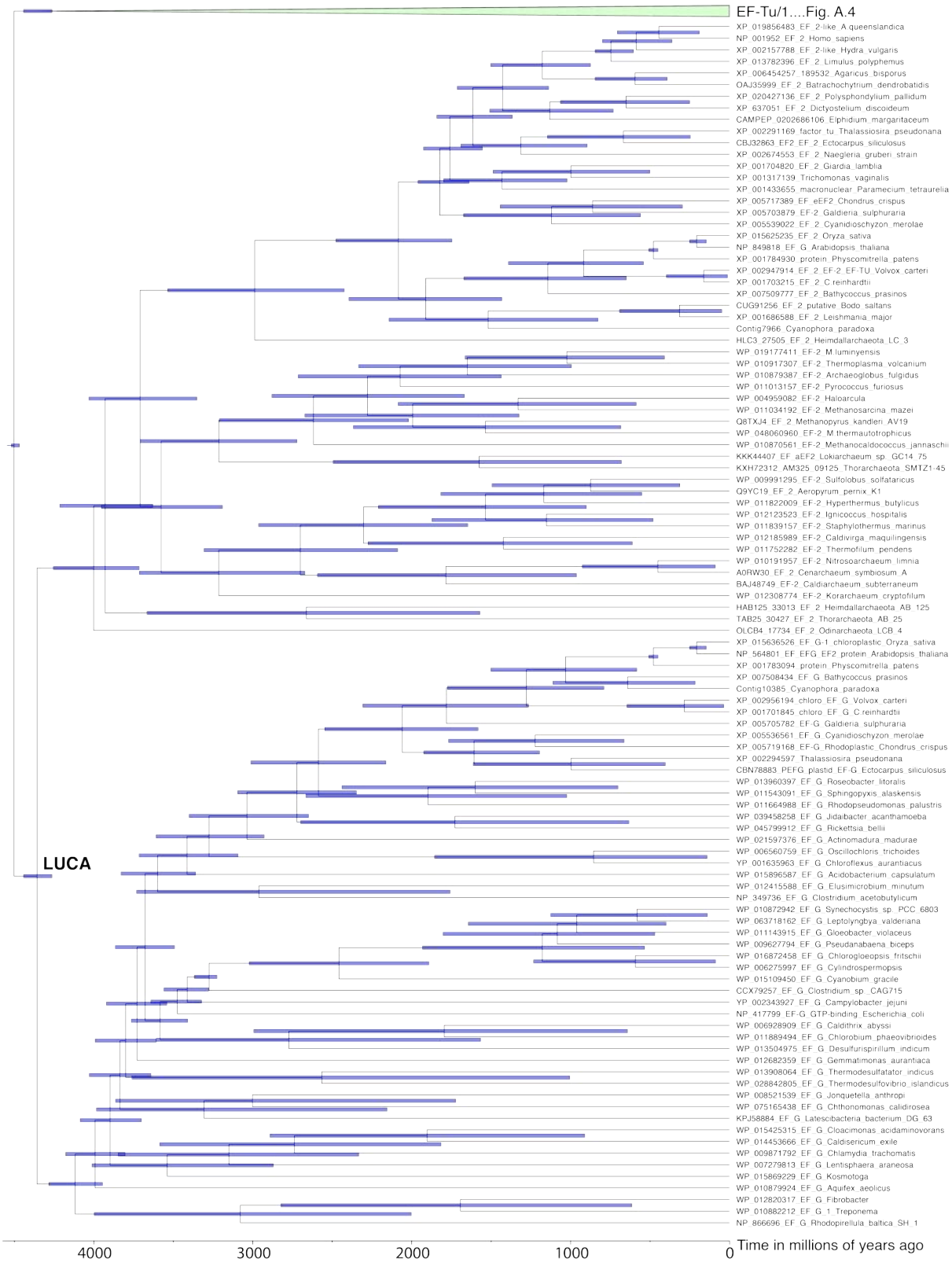


Figure A.3. Divergence time tree of the elongation factor gene family with a focus on the EF-G/2 gene. Blue bars indicate the 95% credible intervals.

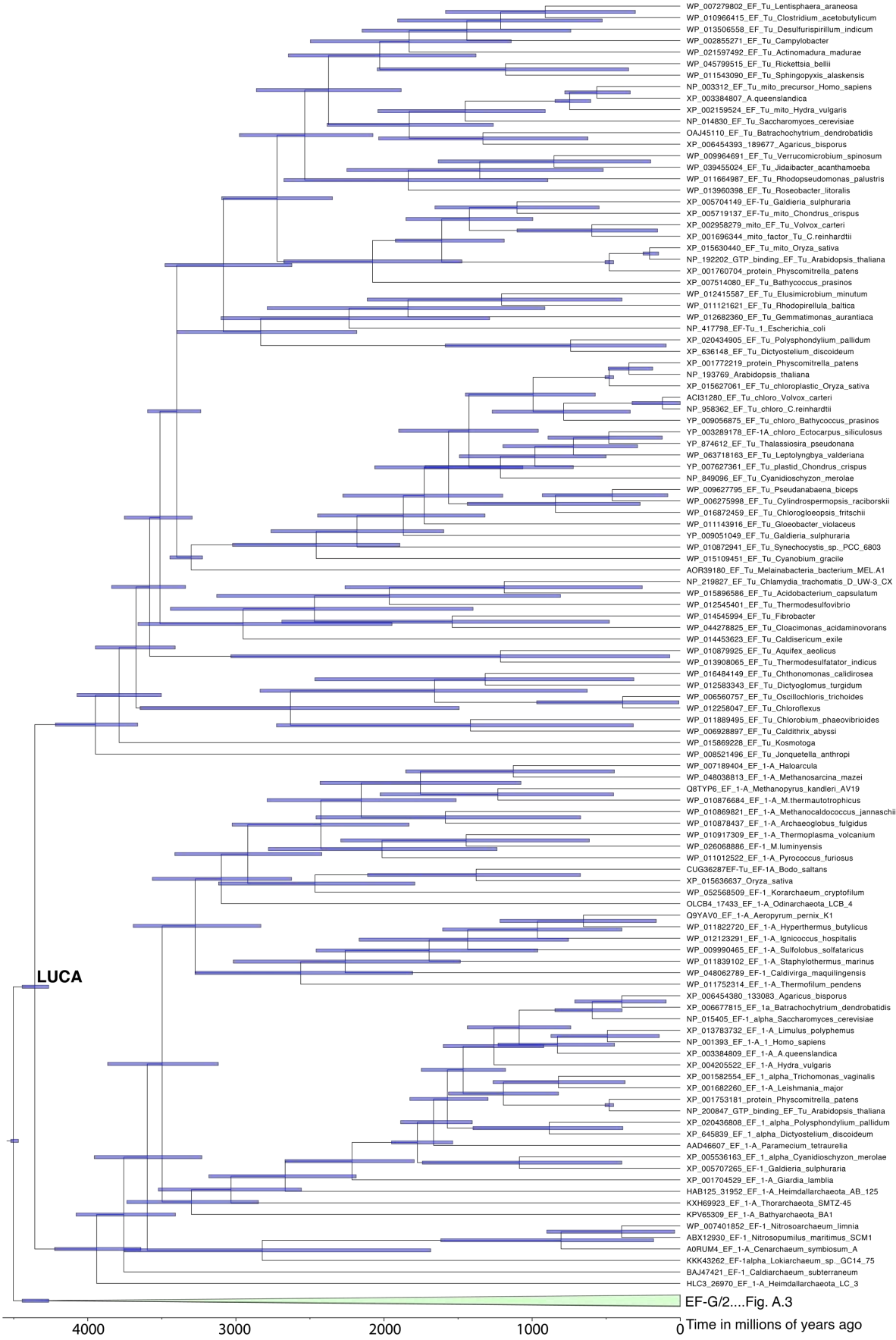


Figure A.4. Divergence time tree of the elongation factor gene family with a focus on the EF-Tu/1 gene. Blue bars indicate the 95% credible intervals.

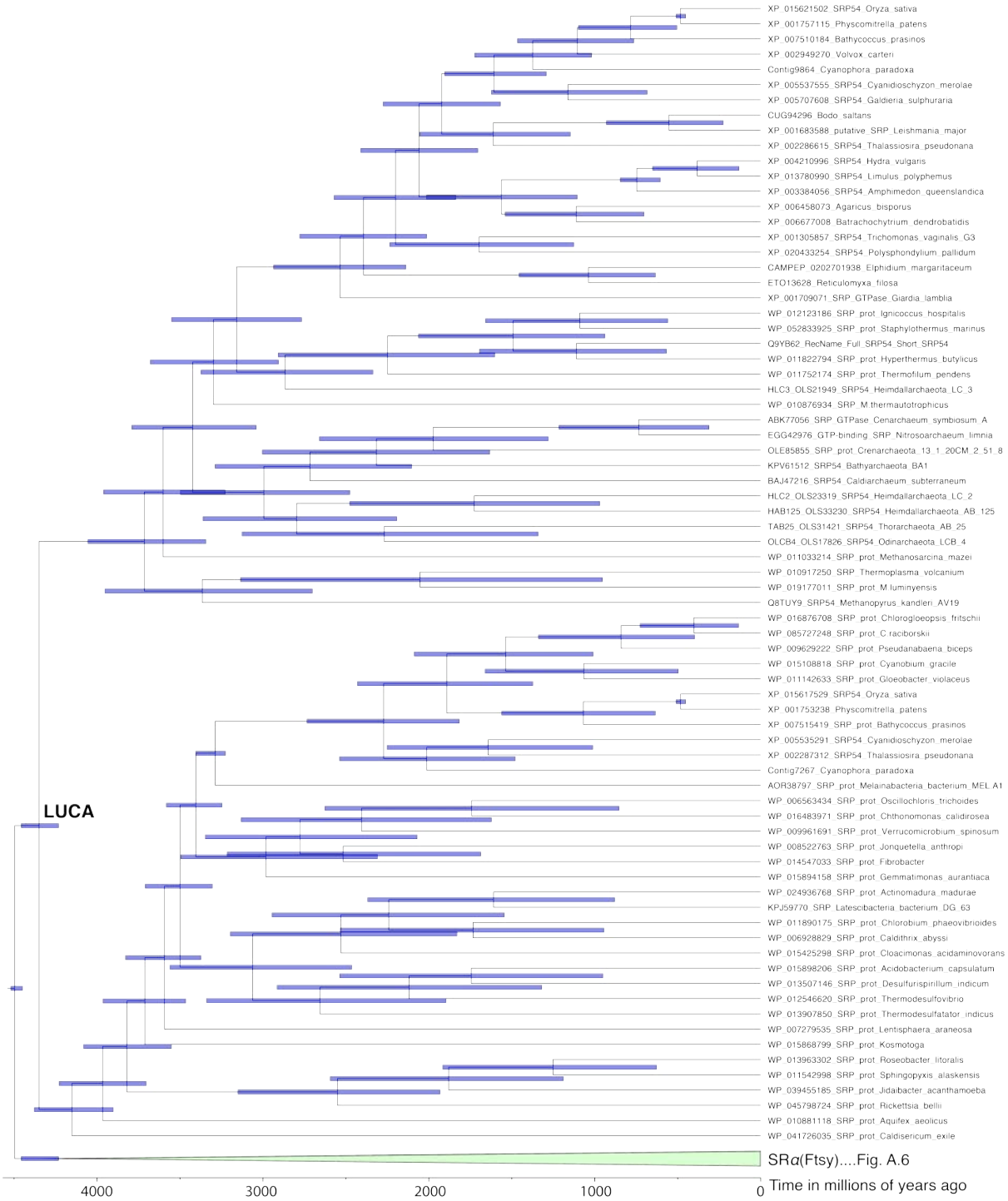


Figure A.5. Divergence time tree of the signal recognition protein gene family with a focus on SRP54(Ffh). Blue bars indicate the 95% credible intervals.

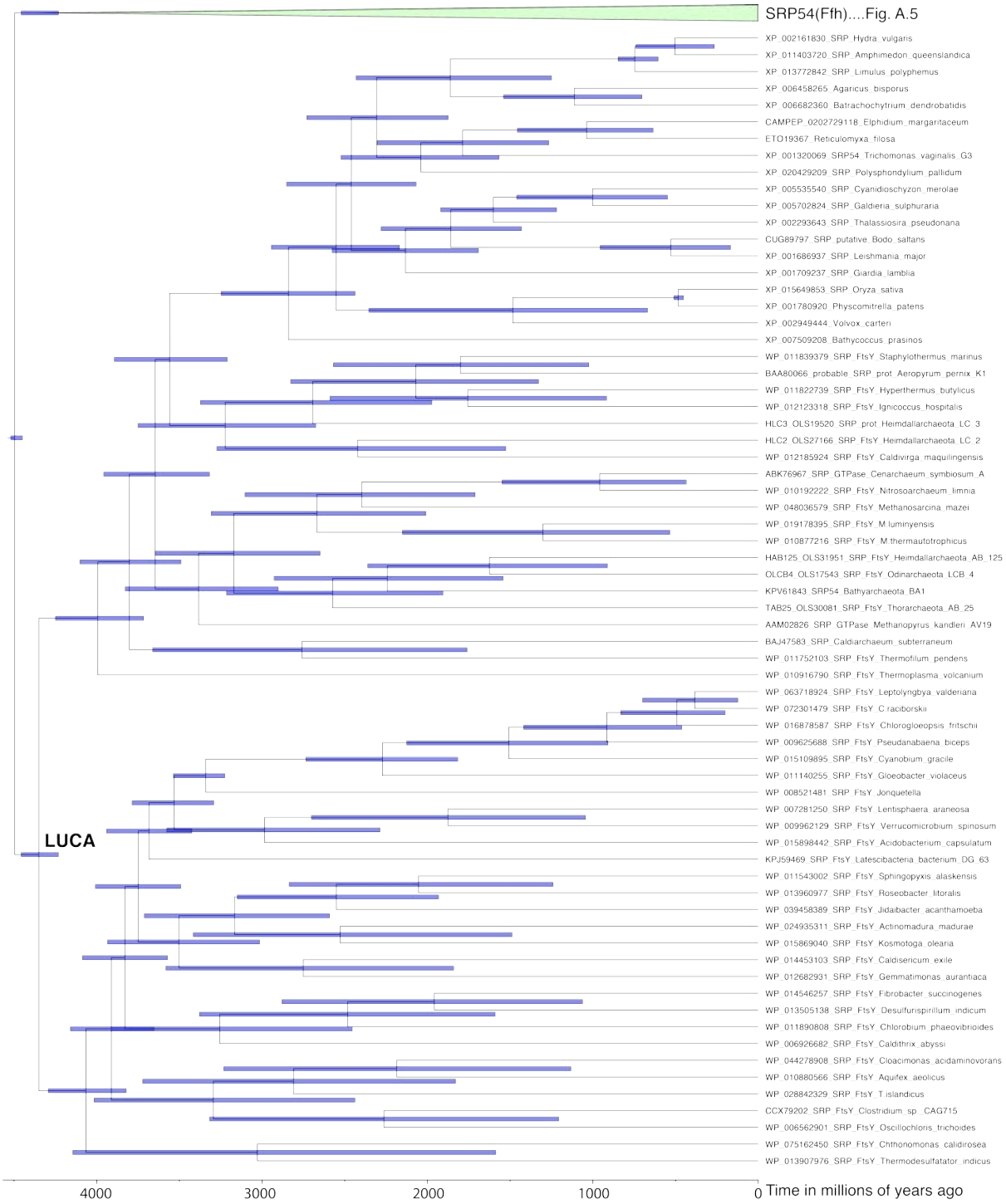


Figure A.6. Divergence time tree of the signal recognition protein gene family with a focus on SRP*a*(Ftsy). Blue bars indicate the 95% credible intervals.

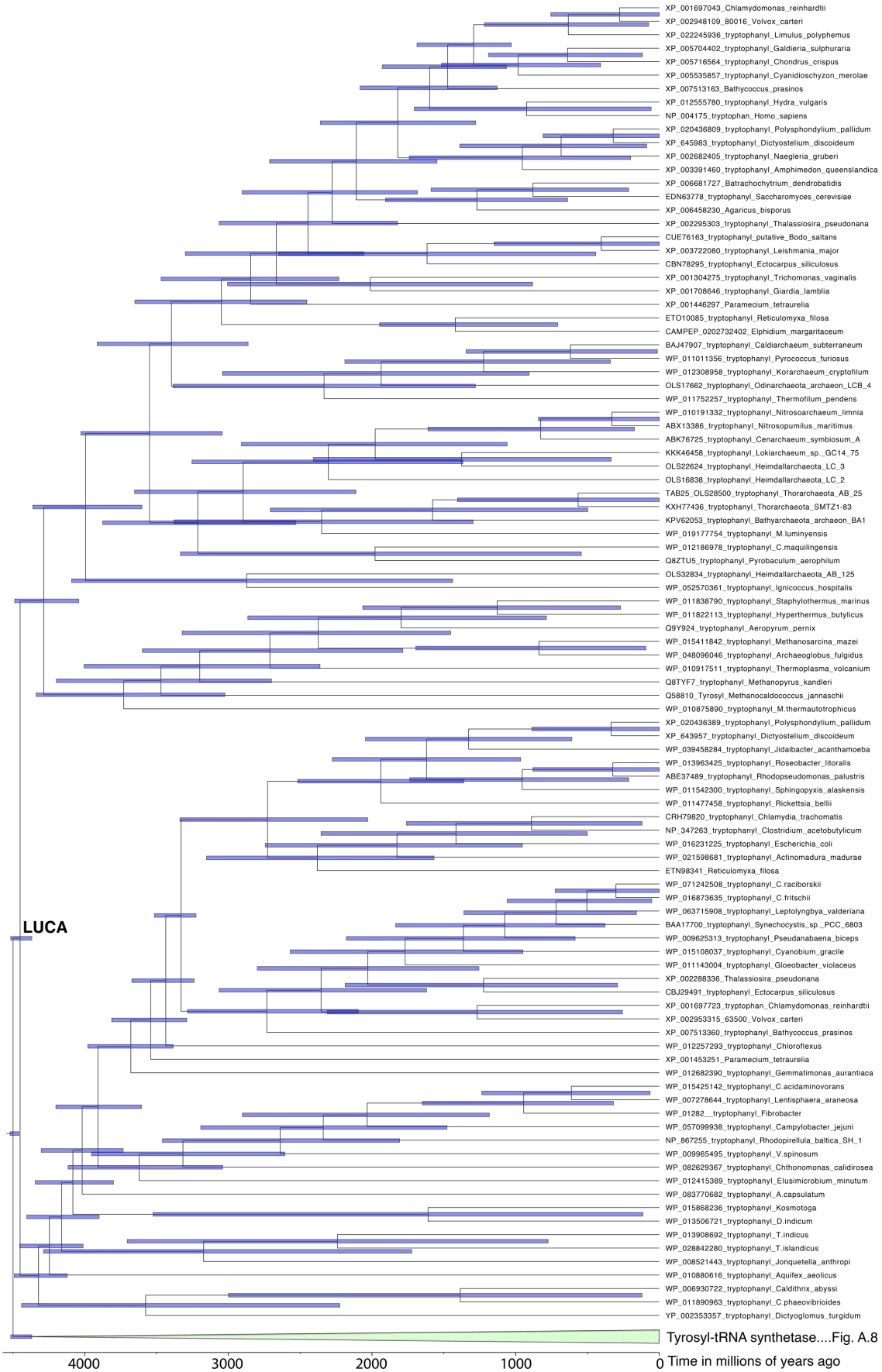


Figure A.7. Divergence time tree of the Tryptophanyl-tRNA and Tyrosyl-tRNA synthetase gene family with a focus on Tryptophanyl-tRNA synthetase. Blue bars indicate the 95% credible intervals.

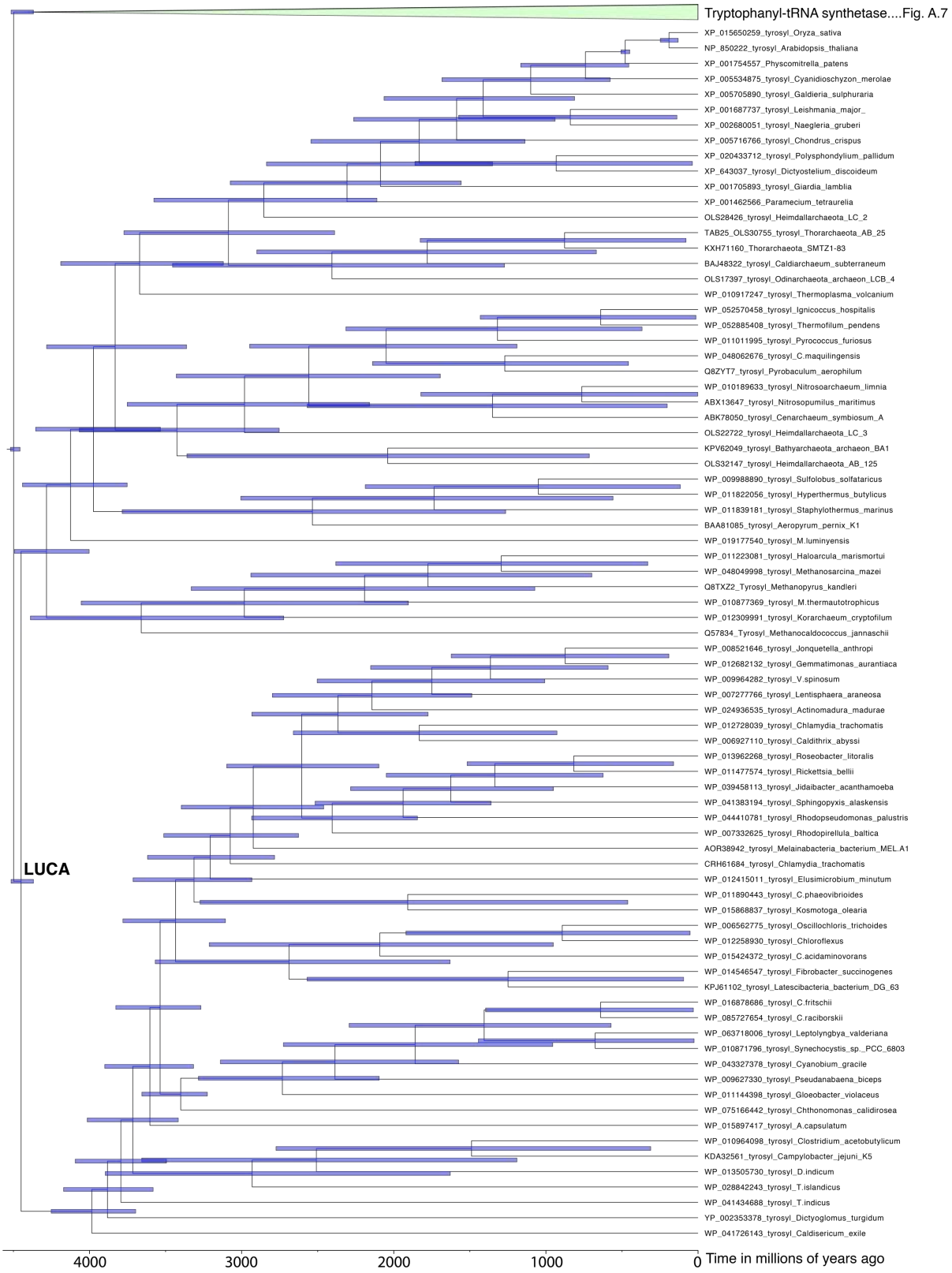


Figure A.8. Divergence time tree of the Tryptophanyl-tRNA and Tyrosyl-tRNA synthetase gene family with a focus on Tyrosyl-tRNA synthetase. Blue bars indicate the 95% credible intervals.

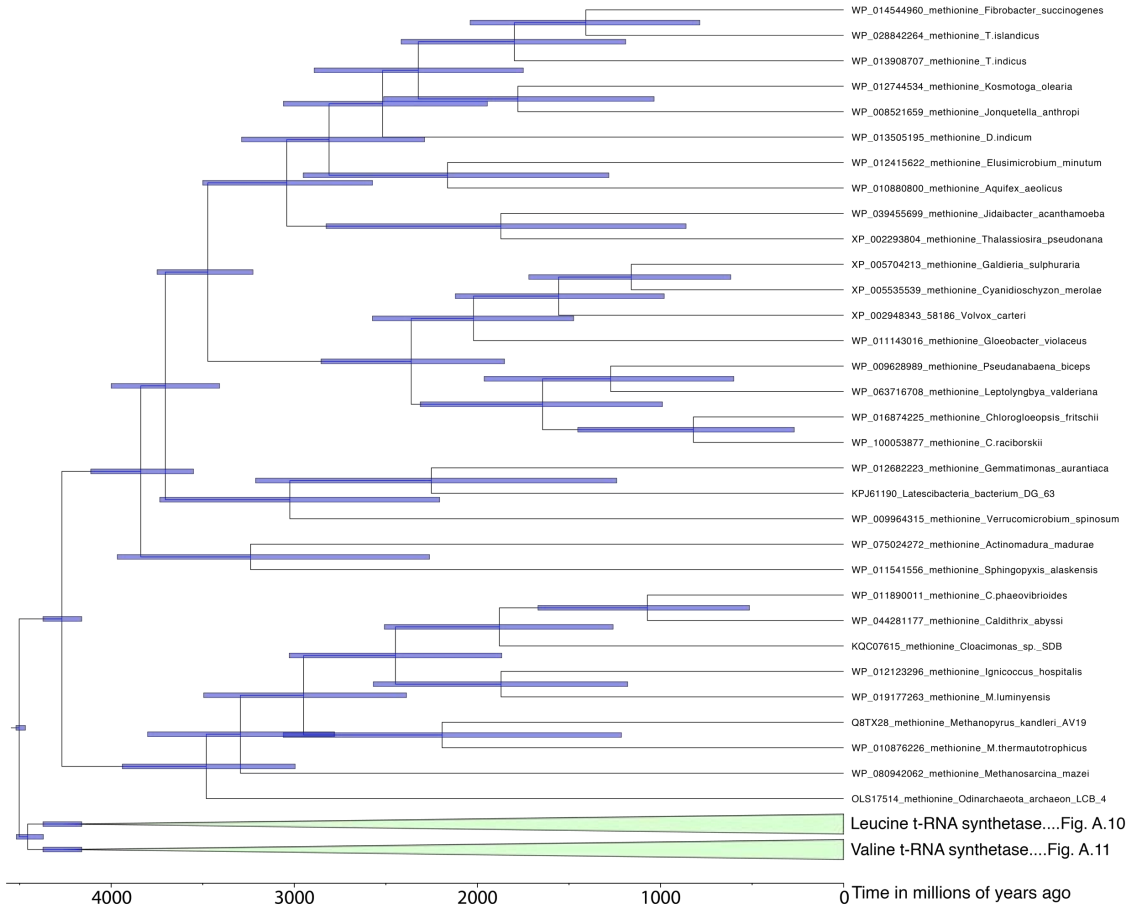


Figure A.9. Divergence time tree of the Methionyl, Leucyl and Valyl-tRNA synthetase gene family with a focus on Methionyl-tRNA synthetase. Blue bars indicate the 95% credible intervals.

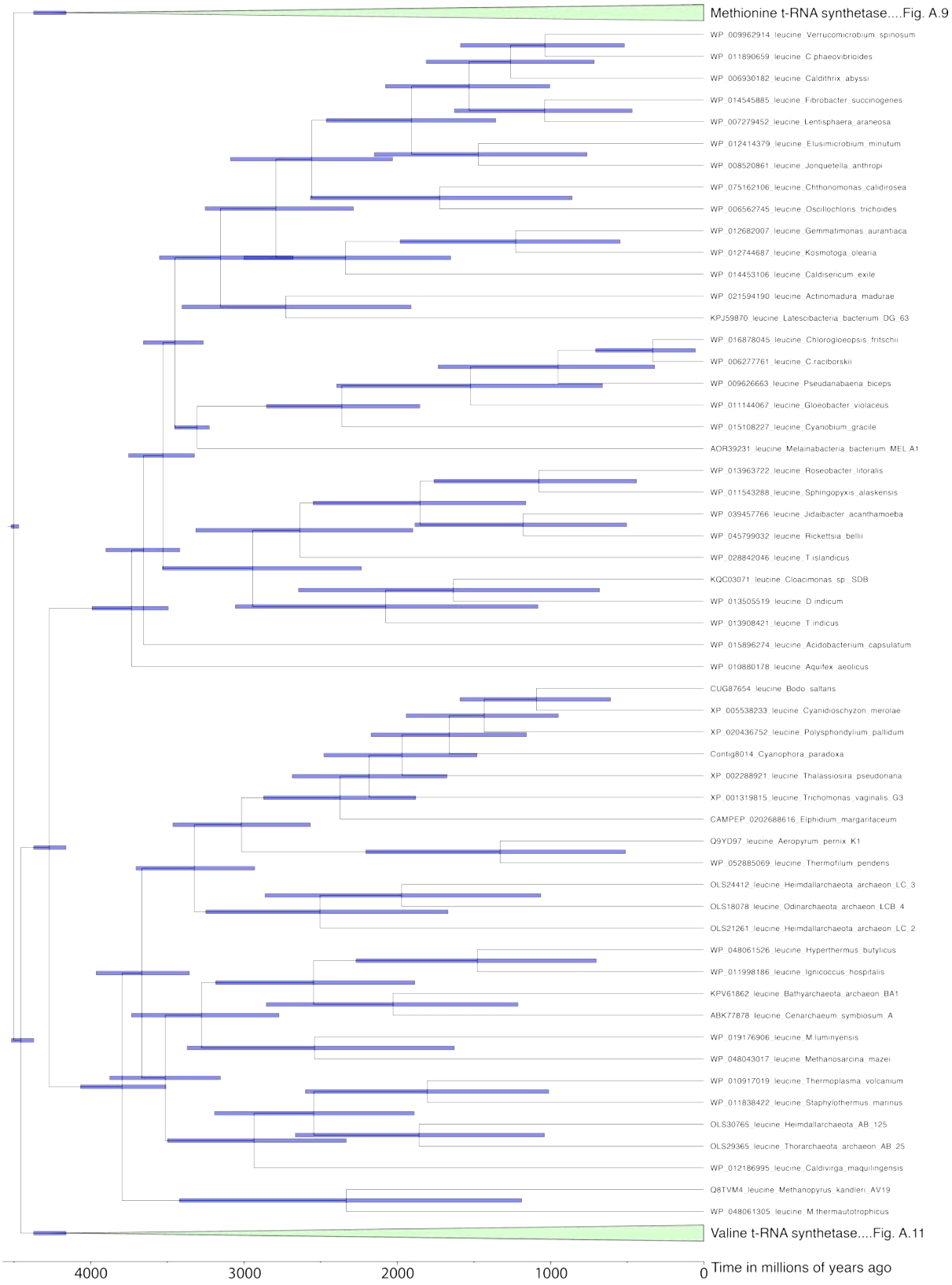


Figure A.10. Divergence time tree of the Methionyl, Leucyl and Valyl-tRNA synthetase gene family with a focus on Leucyl-tRNA synthetase. Blue bars indicate the 95% credible intervals.

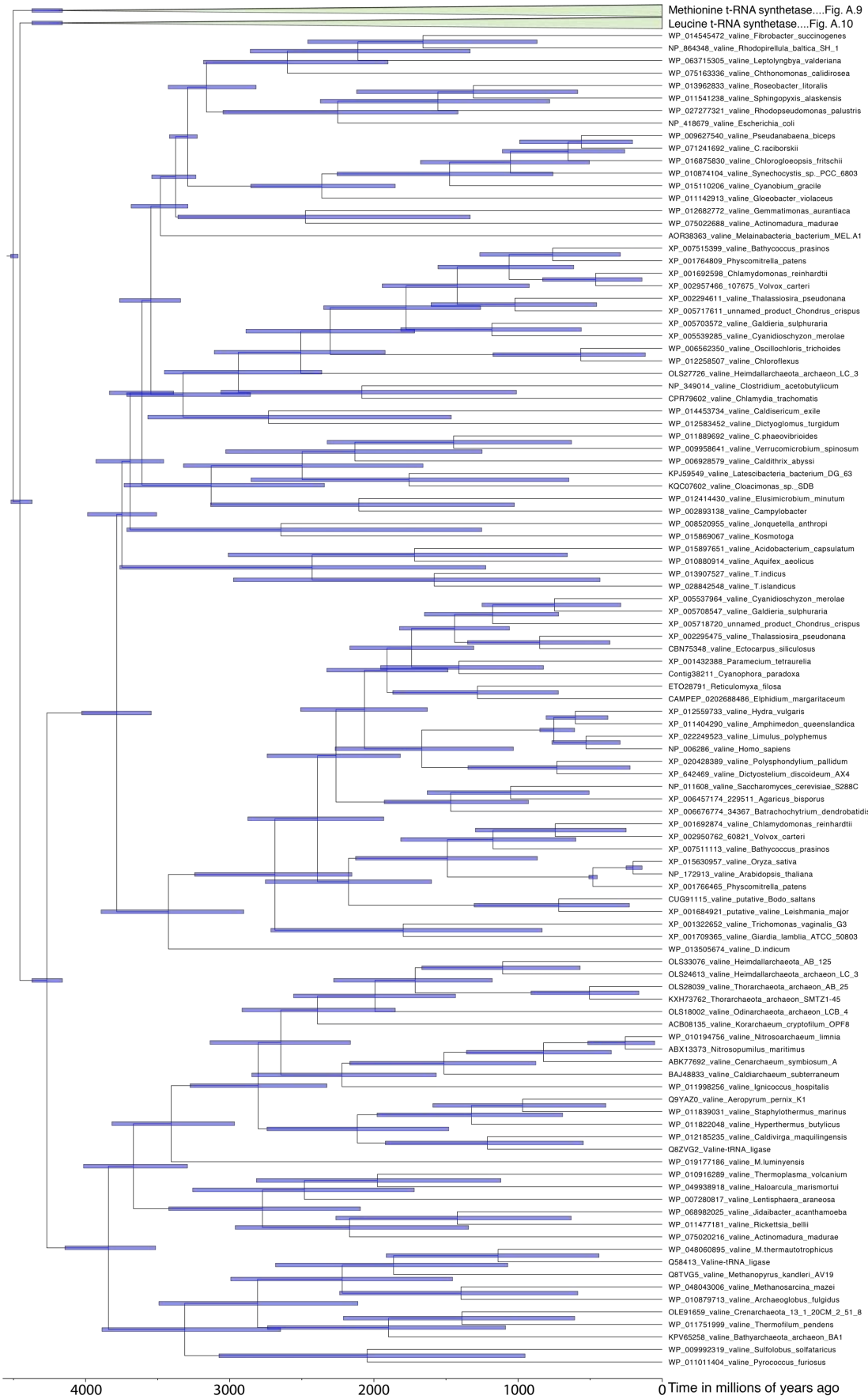


Figure A.11. Divergence time tree of the Methionyl, Leucyl and Valyl-tRNA synthetase gene family with a focus on Valyl-tRNA synthetase. Blue bars indicate the 95% credible intervals.

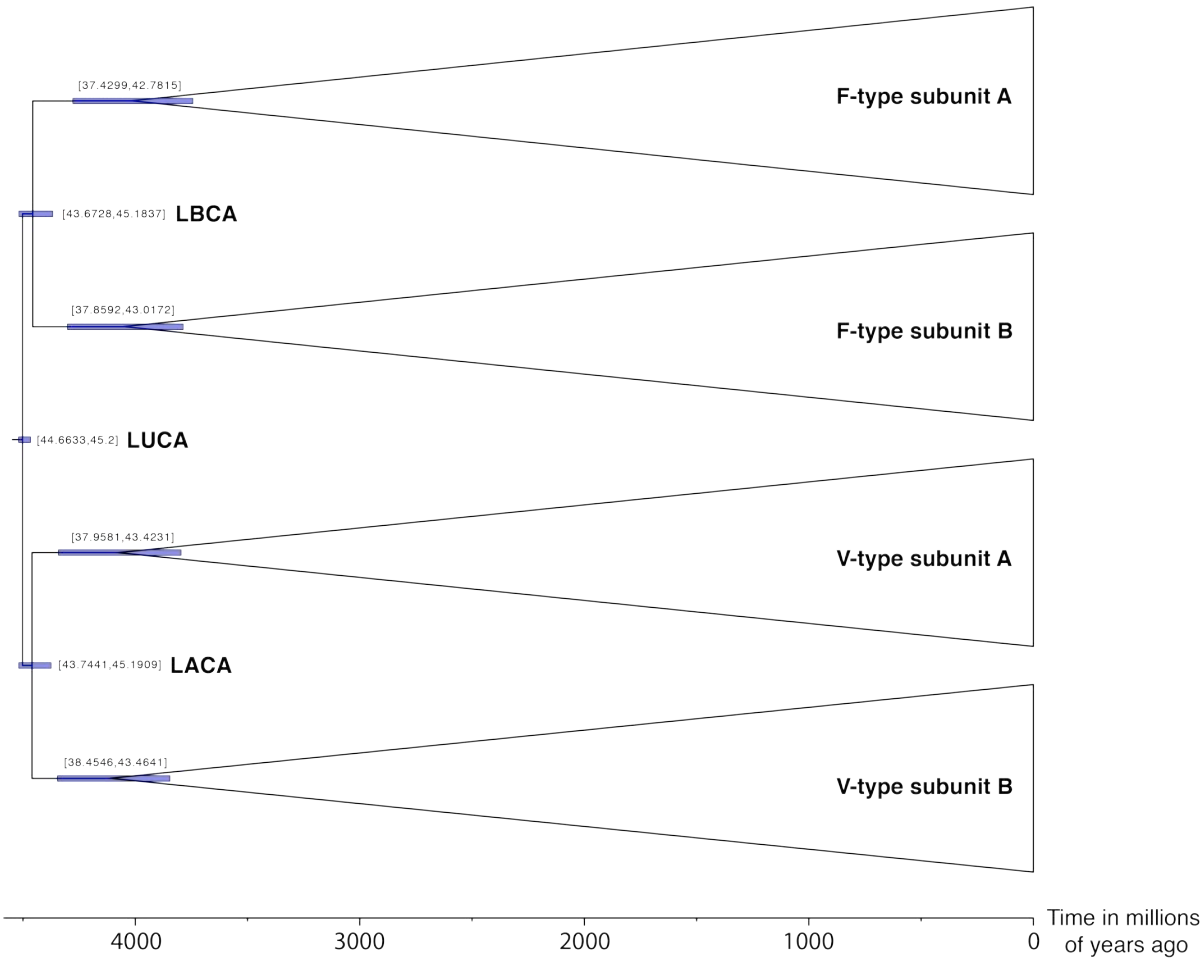


Figure A.12. Divergence time tree of the F-type and V-type ATPase gene family where the F-type subunits and the V-type subunits group together. This means that there is only one last universal common ancestor node. LBCA = last bacterial common ancestor and LACA = last archaeal common ancestor. Blue bars indicate the 95% credible intervals.

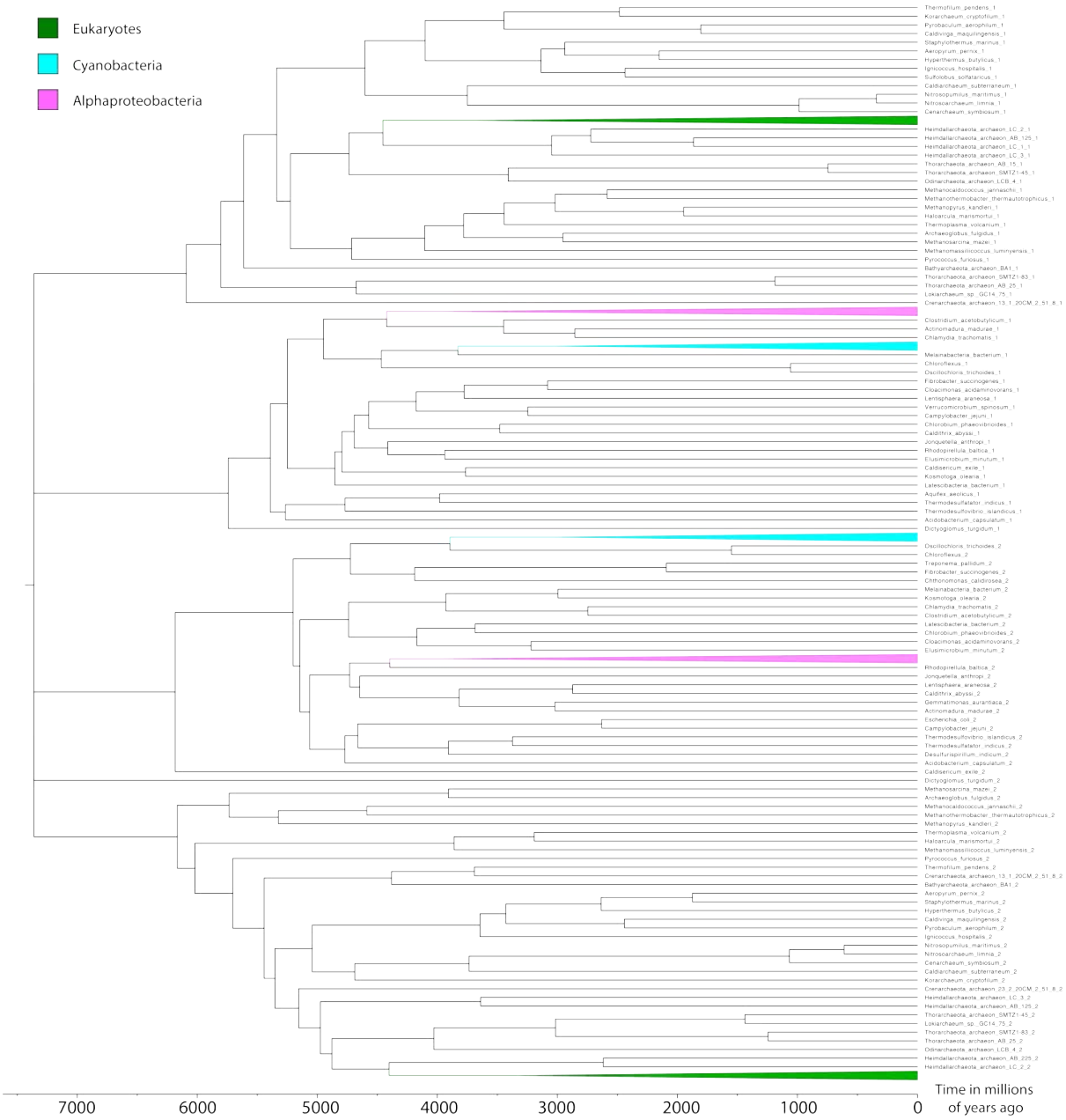


Figure A.13. Divergence time tree produced using a concatenated dataset and cross-bracing. This time tree exceeds both age of the Earth and the moon forming impact (See calibrations section 2.3.1).

Appendix B

Published Articles

nature ecology & evolution

ARTICLES

https://doi.org/10.1038/s41559-018-0644-x

Integrated genomic and fossil evidence illuminates life's early evolution and eukaryote origin

Holly C. Betts[1], Mark N. Puttick[1,2], James W. Clark[1], Tom A. Williams[1,3], Philip C. J. Donoghue[1] and Davide Pisani[1,3]*

Establishing a unified timescale for the early evolution of Earth and life is challenging and mired in controversy because of the paucity of fossil evidence, the difficulty of interpreting it and dispute over the deepest branching relationships in the tree of life. Surprisingly, it remains perhaps the only episode in the history of life where literal interpretations of the fossil record hold sway, revised with every new discovery and reinterpretation. We derive a timescale of life, combining a reappraisal of the fossil material with new molecular clock analyses. We find the last universal common ancestor of cellular life to have predated the end of late heavy bombardment (>3.9 billion years ago (Ga)). The crown clades of the two primary divisions of life, Eubacteria and Archaebacteria, emerged much later (<3.4 Ga), relegating the oldest fossil evidence for life to their stem lineages. The Great Oxidation Event significantly predates the origin of modern Cyanobacteria, indicating that oxygenic photosynthesis evolved within the cyanobacterial stem lineage. Modern eukaryotes do not constitute a primary lineage of life and emerged late in Earth's history (<1.84 Ga), falsifying the hypothesis that the Great Oxidation Event facilitated their radiation. The symbiotic origin of mitochondria at 2.053–1.21 Ga reflects a late origin of the total-group Alphaproteobacteria to which the free living ancestor of mitochondria belonged.

Attempts to investigate the emergence of life and its subsequent evolution have traditionally focused on the fossil record. However, this record, especially when looking at the earliest scions of life, is minimal and interpretation is made harder due to difficulties substantiating relationships within the earliest branching lineages of the tree of life[1,2]. Despite its problematic nature, the fossil record remains the main source of information for the timeline of life's evolution. We attempt to shed light on this early period by presenting a molecular timescale based on the ever-growing collection of genetic data, and explicitly incorporating uncertainty associated with fossil sampling, ages and interpretations[1,3–5].

Calibrations are a crucial component of divergence time estimation. Relative divergence times can be inferred using alternative lines of evidence; for example, horizontal gene transfers[6]. However, an absolute timescale for evolutionary history can only be derived when calibrations are included in the analyses[7,8]. We derived a suite of calibrations, following best practice[4] for the fundamental clades within the tree of life, drawing on multiple lines of evidence, including physical fossils, biomarkers and isotope geochemistry[2]. Two key calibrations, for the last universal common ancestor (LUCA) and the oldest total-group eukaryotes, constrain the whole tree by setting a maximum on the root, while also informing the timing of divergence of eukaryotes within Archaea[9,10]. Putative records for life extend back to the Eoarchaean, including microfossils[11,12], stromatolites[13] and isotope data[14,15] from the ~3.8 billion years ago (Ga) Isua Greenstone Belt (Greenland). However, these records have been contested[16–18]. Microfossils from the ~3.4 Ga Strelley Pool Formation, Australia, are the oldest conclusive evidence to constrain the age of LUCA[19]. The fossils, many of which are arranged in chains of cells, have been shown, through nanoscale imaging and Raman spectroscopy, to exhibit a complex morphology with a central, usually hollow, lenticular body and a wall that is either smooth or in some cases reticulated; these features are beyond the scope of pseudofossils[2]. The Strelley Pool Formation also contains other microfossils[20–22], in association with both distinct $\partial^{13}C_{org}$ and $\partial^{13}C_{inorg}$[23] and pyrite indicative of sulfur metabolisms[24], along with stromatolites that exhibit biological structure[25]. Overall, these data allow us to confidently use the Strelley Pool Biota as the oldest, undisputable, record of life. For a maximum constraint on the age of LUCA, we considered the youngest event on Earth that life could not have survived. Conventionally, this is taken as the end of the episode of late heavy bombardment, but modelling has shown that this would not have been violent enough for planet sterilization[26]. However, the last formative stage of Earth's formation—the Moon-forming impact—melted and sterilized the planet. The oldest fossil remains that can be ascribed to crown Eukaryota are ~1.1 Ga *Bangiomorpha pubescens*[27,28], which can be confidently assigned to the red algal total group (Rhodophyta). Older fossil remains from the >1.561 Ga Chitrakoot Formation have been tentatively interpreted as red algae[29]; however, current knowledge of their morphology does not allow for an unequivocal assignment to crown Archaeplastida. The oldest fossil remains that can be ascribed with certainty to total-group Eukaryota are acritarchs from the >1.6191 Ga Changcheng Formation, North China[30], which are discriminated from prokaryotes by their large size (40–250 µm) and complex wall structure, including striations, longitudinal ruptures and a trilaminar organization. However, these structures do not indicate membership of any specific crown eukaryote clade, only allowing us to use these records to minimally constrain the timing of divergence between the Eukaryota and their archaebacterial sister lineage, Asgardarchaeota[9,10,31]. As there is no other evidence to maximally constrain the time of divergence between Eukaryota and Asgardarchaeota, we used the same maximum placed on LUCA; that is, the Moon-forming impact. These key time constraints were combined with nine others (see Supplementary Information) to calibrate a timescale of life estimated from a dataset of 29 highly

[1]School of Earth Sciences, University of Bristol, Bristol, UK. [2]Milner Centre for Evolution, Department of Biology and Biochemistry, University of Bath, Bath, UK. [3]School of Biological Sciences, University of Bristol, Bristol, UK. *e-mail: davide.pisani@bristol.ac.uk

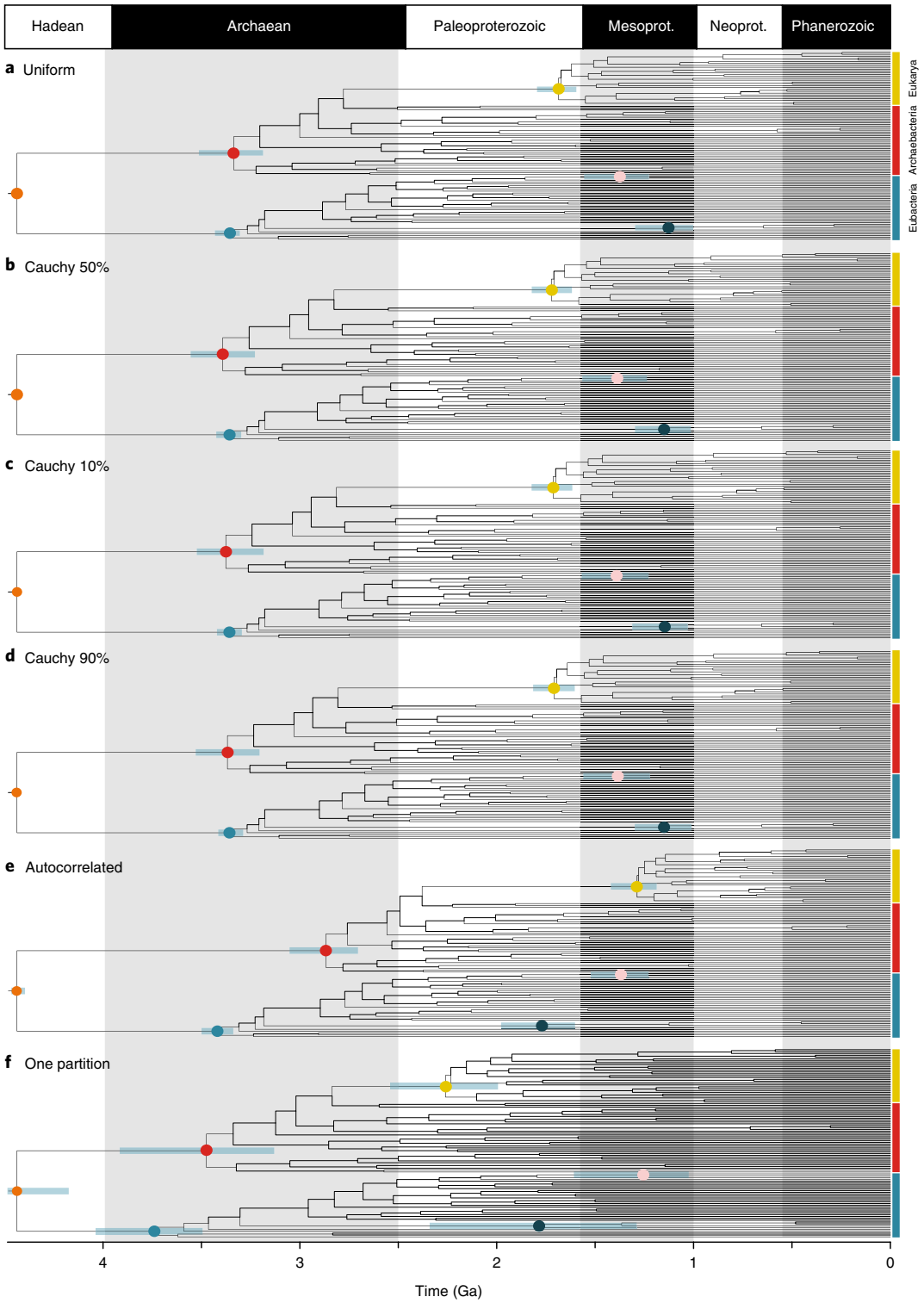


Fig. 1 | Posterior time estimates under different parameters. a, Posterior time estimates when using a uniform calibration density prior distribution, reflecting a lack of information about the divergence time relative to the fossil constraint. **b**, Cauchy 50% maximum calibration density prior distribution, reflecting a view that the divergence date should fall between the constraints. **c**, Cauchy 10% maximum calibration density prior distribution, reflecting a view that the fossil prior is a good approximation of the divergence date. **d**, Cauchy 90% maximum calibration density prior distribution, reflecting a view that the fossil prior is a poor approximation of the divergence date, all with an uncorrelated clock model. **e**,**f**, Posterior age estimates when using a Cauchy 50% maximum calibration density prior distribution with an autocorrelated clock model (**e**) and with an uncorrelated clock model and a single partition scheme (**f**). All molecular clock analyses converged well. The coloured dots highlight specific nodes, with their respective confidence intervals displayed light blue bars (orange, LUCA; red, crown Archaeabacteria; blue, crown Eubacteria; yellow, crown Eukaryota; pink, alphaproterobacteria; dark blue, cyanobacteria). This figure illustrates how divergence times change as alternative approaches to modelling calibrations and the process of molecular evolution are implemented. Divergence estimates from **f** and their credibility intervals could be rejected based on an AIC test. The other results (**a**–**e**) cannot be rejected. Mesoprot., Mezoproterozoic; Neoprot., Neoproterozoic.

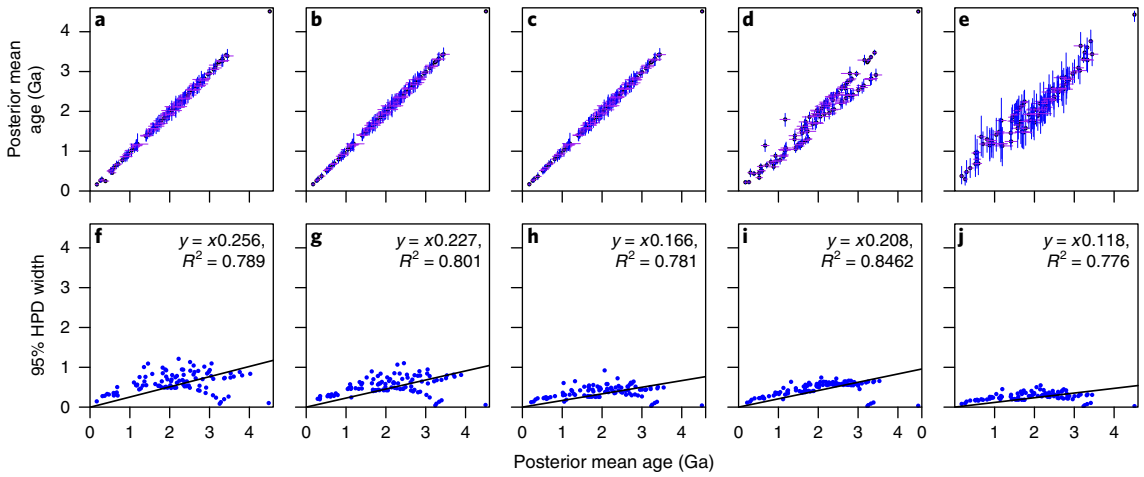


Fig. 2 | Changes in divergence times (Ga) that result from applying alternative parameters. a, Cauchy 50% maximum calibration density prior distribution versus uniform calibration density prior distribution. **b**, Cauchy 50% maximum calibration density prior distribution versus Cauchy 10% maximum calibration density prior distribution. **c**, Cauchy 50% maximum calibration density prior distribution versus Cauchy 90% maximum calibration density prior distribution. **d**, Cauchy 50% maximum calibration density prior distribution uncorrelated clock model versus Cauchy 50% maximum calibration density prior distribution autocorrelated clock model. **e**, Cauchy 50% maximum calibration density prior distribution in both cases for the 29-partition scheme versus the 1-partition scheme. **f**–**j**, Results of adding additional genes as infinite sites plots: 5-gene dataset (**f**); 10-gene dataset (**g**); 15-gene dataset (**h**); 20-gene dataset (**i**); 29-gene dataset (**j**). Blue dots denote node dates. HPD, highest posterior density.

conserved, mainly ribosomal, universally distributed proteins (see Supplementary Information) using a relaxed molecular clock modelled in a Bayesian framework.

Results

Analytical choices can deeply affect molecular clock posterior age estimates[32] and we explored a range of prior probability distributions to model our fossil calibrations and estimate conservative credibility intervals for our divergence times. Initially, we applied a hard maximum of 4.52 Ga (the age of the Moon-forming impact) to the root of our tree and used uniform age priors (reflecting agnosticism about divergence timing relative to constraints) to the other fossil calibrations (Fig. 1a). These analyses assumed an uncorrelated molecular clock model and produced the amino acid substitution processes using optimal gene-specific substitution models. Subsequently, we explored the impact of using calibration protocols based on non-uniform age priors. First, we implemented a truncated Cauchy distribution with the mode located halfway between the minimum and maximum bounds, reflecting a prior view that true divergence times should fall between the minimum and maximum calibration points (Fig. 1b). In two subsequent analyses we applied a skewed Cauchy distribution such that the mode shifted towards the minimum or the maximum constraint, reflecting prior views that the fossils used to calibrate the tree are either very good (Fig. 1c) or very poor (Fig. 1d) proxies of the true divergence times. Our results proved robust to the use of different calibration strategies, only identifying some variability in the size of the recovered credibility intervals (Fig. 2a–c).

We explored the impact of different strategies for modelling both the molecular clock (Fig. 1e) and the amino acid substitution process (Fig. 1f). Only minimal differences in posterior ages were found between analyses using an uncorrelated or autocorrelated clock (Fig. 2d). Consistently, Bayesian cross-validation indicated that the two models do not differ significantly in their fit to the data (cross-validation score = 0.7 ± 2.96816 in favour of the uncorrelated clock). In contrast, using a single substitution model across the 29 genes or using an optimal set of gene-specific substitution models inferred using PartitionFinder[33] resulted in very different age estimates (Figs. 1f and 2e). Using a single substitution model recovered larger credibility intervals (Fig. 2e) with a more homogeneous distribution of branch lengths across the tree, and older divergence times (compare Fig. 1f and Fig. 1a–d). An Akaike information criterion (AIC) test indicated that the partitioned model provides a significantly better fit to the data (AIC score = 565.21 in favour of 29 gene-specific models), allowing the rejection of the divergence times obtained with a single substitution model. As expected, divergence times estimated from individual genes were much less precise, although posterior age estimates overlap well (Supplementary Section 4.1). This indicates that the genes comprising our dataset encode a congruent signal and the timescale inferred from the combined analysis is not biased by single gene outliers. Furthermore, their combination improves the precision of the clade age estimates (Fig. 2f–j), which are clearly informed by the data (Supplementary Section 4.2). We tested the effect of taxonomic sampling by doubling the number of cyanobacteria and alphaproteobacteria in our dataset. We then explored the effect of phylogenetic uncertainty by dating a tree compatible with Woese's three-domains hypothesis[34] and by dating all 15 trees in the 95% credible set of trees from our phylogenetic analysis (Supplementary Sections 4.3 and 4.4). Further analyses that used co-estimation of tree and topology (Supplementary Section 4.5)[35] did not reach convergence (Supplementary Section 4.6), but the results recovered were congruent with those obtained from well-converged analyses (Supplementary Section 4.4) where topology and time were inferred sequentially (see the caption of Supplementary Section 4.5 for a discussion). Overall, the outcome of these experiments demonstrates that our original results are robust to topological uncertainty and the use of differential taxonomic sampling (Supplementary Sections 4.3–4.5).

It is not possible to discriminate between the competing calibration strategies that reflect different interpretations of the fossil record. Similarly, our model selection test indicated that the autocorrelated and independent-rates clock models fit the data equally well. Thus, in establishing an accurate timescale of life, we integrated over the uncertainties associated with the results from all these analyses (Fig. 3). The joint 95% credibility intervals reject a post-late heavy bombardment (~3,900 million years ago (Ma)[36] emergence of LUCA (4,519–4,477 Ma). The crown clades of the primary divisions of life, Archaebacteria and Eubacteria emerged over one billion years after LUCA in the Mesoarchaean–Neoarchaean. The earliest conclusive evidence of cellular life (Strelley Pool Formation,

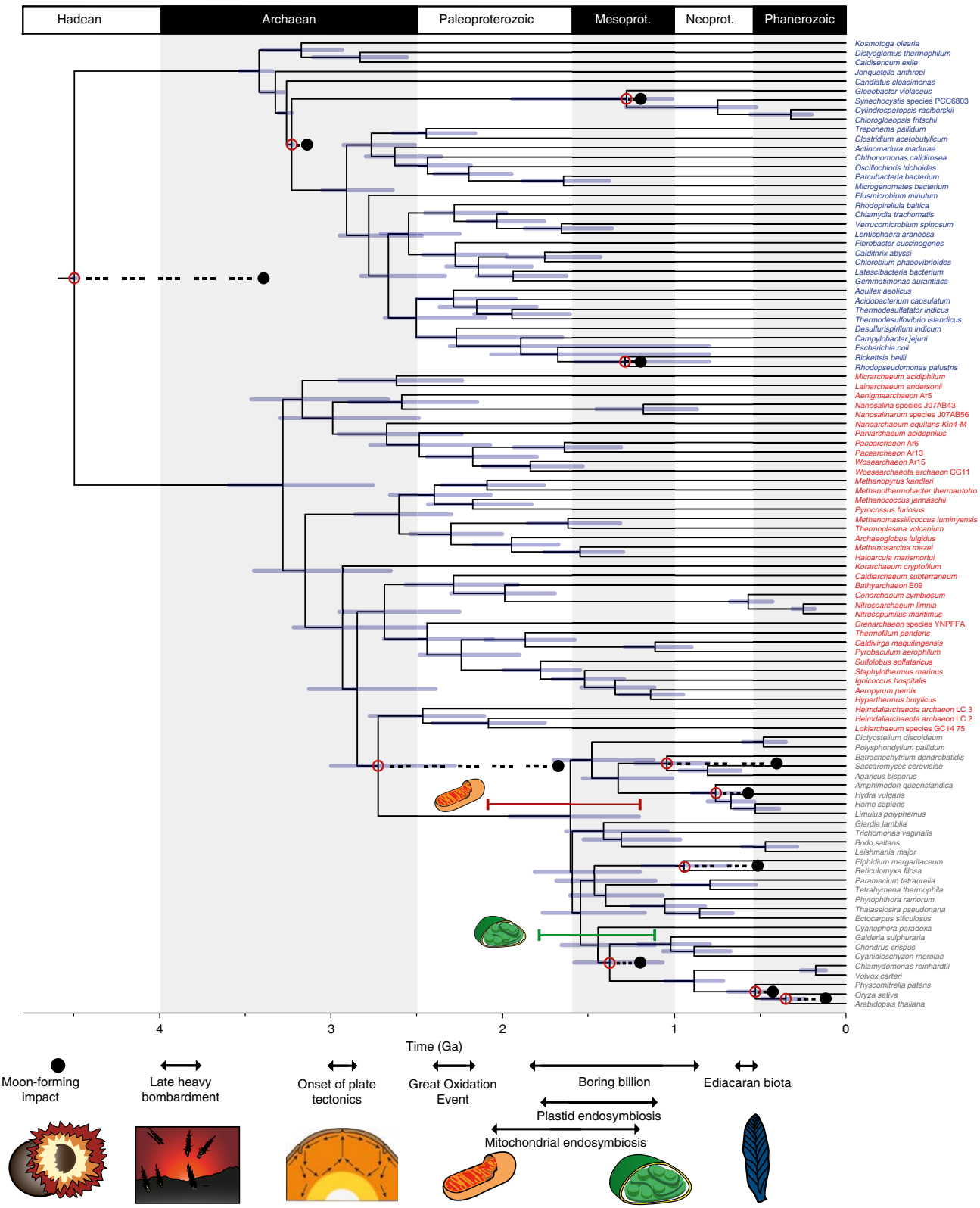


Fig. 3 | A tree combining uncertainties from approaches using uncorrelated and autocorrelated clock models and different calibration density distributions. Tip labels are shown for Eukaryota (grey), Archaeabacteria (red) and Eubacteria (blue). The purple bars denote the credible intervals for each node. Red dots highlight calibrated nodes, and corresponding black dots highlight the age of the minimum bound of its corresponding calibration. The phylogenetic relationships of the mitochondrion within Alphaproteobacteria are still debated[56,74–76], and it is unclear whether the free-living ancestor of the mitochondrion was a crown or stem representative of this group. The red bar above the crown eukaryote node denotes the time period during which the mitochondrial endosymbiosis may have occurred. The green bar denotes the time during which the plastid endosymbiosis may have occurred. Important events in Earth and life history are indicated along the base of the figure. Mesoprot., Mezoproterozoic; Neoprot., Neoproterozoic.

Australia[2]) falls within the 95% credibility intervals for the ages of the last common ancestors of both clades, indicating that these fossils might belong to one of the two living prokaryotic lineages.

Discussion

Methanogenesis is classically associated with Euryarchaeota. Our estimate for the age of crown Euryarchaeota (2,881–2,425 Ma) is consistent with carbon isotope excursions indicating the presence of methanogens by 2 Ga[37], but is substantially younger than the earliest possible evidence of biogenic methane in the geochemical record at ~3.5 Ga[38,39]. If the geochemical evidence is correct, our timescale implies that methanogenesis predated the origin of Euryarchaeota. This hypothesis would be consistent with recent environmental genomic surveys indicating that other archaeal lineages may also be capable of methane metabolism[40] or methanogenesis[41], and that metabolisms using the Wood–Ljungdahl pathway to fix carbon minimally evolved in stem archaebacteria[42,43] and might have been a characteristic of LUCA[43–45].

The Great Oxidation Event (GOE; ~2.4 Ga) was perhaps the most significant episode in the Proterozoic[46], fundamentally changing the chemistry of Earth's atmosphere and oceans, and probably altering temperature. It has been causally associated with the evolution of Cyanobacteria, as a consequence of their oxygen release[28,47], and implicated as an extrinsic driver of eukaryotic evolution[48]. Our timescale indicates that crown Cyanobacteria and crown Eukaryota significantly postdate the GOE. Crown Cyanobacteria diverged 1,947–1,023 Ma, precluding the possibility that oxygenic photosynthesis emerged in the cyanobacterial crown ancestor. However, the Cyanobacteria separated from other eubacterial lineages (Fig. 3), including the non-photosynthetic sister group of the Cyanobacteria (Melanibacteria; Supplementary Section 4.3) in the Archaean, before the GOE, consistent with the view that oxygenic photosynthesis evolved along the cyanobacterial stem[49], and compatible with a causal role of the total-group Cyanobacteria in the GOE.

Crown Eukaryota diverged considerably after both the Eukaryota–Asgardarchaeota split and the GOE, in the middle Proterozoic (1,842–1,210 Ma). Our study strongly rejects the idea that eukaryotes might be as old as, or older than, prokaryotes[50], and agrees with a number of other studies that date the last eukaryote common ancestor (LECA) to the Proterozoic (~1,866–1,679 Ma)[51–53]. Within eukaryotes, the main extant clades emerged by the middle Proterozoic, including Opisthokonta (~1,707–1,125 Ma), Archaeplastida (~1,667–1,118 Ma) and SAR (stramenopiles (heterokonts), alveolates and Rhizaria; ~1,645–1,115 Ma). The symbiotic origin of the plastid occurred among stem archaeplastids (~1,774–1,118 Ma), and our 95% credibility interval for the origin of the plastid overlap with the results of other recent studies[28,50,54]. The relatively long stem lineage subtending LECA is intriguing. It is found using both uncorrelated and autocorrelated clock models (Figs. 1e and 2d), and disappears only if a poorly fitting single substitution model is used (Figs. 1f and 2e), suggesting that it is not a modelling artefact. Analyses excluding the hitherto unknown immediate living relatives of Eukaryota[9,31], Asgardarchaeota, had no significant impact on the span of the eukaryote stem lineage, suggesting that its length is robust to taxon sampling (Supplementary Section 4.7).

Our timescale for eukaryogenesis rejects the hypothesis of an inextricable link between the GOE and the origin of eukaryotes[48]. Competing hypotheses for eukaryogenesis hinge on the early versus late acquisition of mitochondria relative to other key eukaryote characters[55–59]. Absolute divergence times cannot discriminate between these hypotheses. However, as the only proposed evidence in support of the mitochondria late[57] hypothesis have been shown to be artefactual[58], the similar age estimates for Alphaproteobacteria and LECA at this stage are most conservatively interpreted as indicating that the process of mitochondrial symbiosis underpinned a rapid process of eukaryogenesis. This process involved a large transfer of genes from the genome of the alphaproteobacterial symbiont to that of the archaeal host[59,60], as predicated on metabolism[55,61].

The search for the earliest fossil evidence of life on Earth has created more heat than light. Although the fossil record remains integral to establishing a timescale for the Tree of Life, it is not sufficient in and of itself. Our integrative molecular timescale encompasses the uncertainty associated with fossil, geological and molecular evidence, as well its modelling, allowing it to serve as a solid foundation for testing evolutionary hypotheses in deep time for clades that do not have a credible fossil record.

Methods

Dataset collation and phylogenetic analysis. The dataset consists of 102 species and 29 universally distributed, protein-coding genes (see Supplementary Information). All our data and scripts are available at https://bitbucket.org/bzxdp/betts_et_al_2017. Proteomes were downloaded from GenBank[62] and putative orthologues were identified using BLAST[63]. The top hits were compiled and aligned into gene-specific files in MUSCLE[64] and trimmed to remove poorly aligned sites using Trimal[65]. To minimize the possible inclusion of paralogues and laterally transferred genes, we generated gene trees (under CAT-GTR + G) in PhyloBayes[66] and excluded sequences when the tree topology suggested that they might have been paralogues. The sequences were then concatenated into a supermatrix using FASconCAT[67], and phylogenetic analyses were performed using PhyloBayes[66]. The superalignment was initially analysed under both GTR + G and CAT-GTR + G[68]. RogueNaRok[69] was used to identify rogue taxa, and analyses were repeated (under both GTR + G and CAT-GTR + G) after unstable taxa were excluded. One final analysis was performed that included only the eukaryotic sequences in our dataset (under CAT-GTR + G). For all PhyloBayes analyses, convergence was tested in PhyloBayes using BPCOMP and TRACECOMP.

Calibrations. In total, we used 11 calibrations spread throughout the tree but mainly found within the Eukaryotes as this group has the best fossil record. Calibration choice was carried out conservatively using coherent criteria[4]. Full details of each calibration used can be found in the Supplementary Information.

MCMCTree analysis. For our clock analyses, we used a constraint tree based on our CAT-GTR + G and GTR + G trees (Supplementary Sections 3.2, 3.3 and 4; see the results of phylogenetic analyses in the Supplementary Information for details). The complete phylogeny was rooted to separate Eubacteria from the other lineages (that is, Archaeabacteria and Eukayota). To select the amino acid model to be used in our molecular clock analyses, we used PartitionFinder version 1.1.1 (ref. [33]). Divergence time estimation was carried out using the approximate likelihood calculation in MCMCTree version 4.9 (ref. [70]). We set four different calibration density distributions: uniform, skewed towards the minimum, skewed towards the maximum and midway between these two dates. For this, we used the Uniform and Cauchy models within MCMCTree, which can be set to place the maximum probability of the node falling in a certain space between the calibrations. The values for these were first produced using MCMCTreeR (https://github.com/PuttickMacroevolution/MCMCTreeR) code in R[71]. We investigated two strategies to model amino acid sequence evolution: a single WAG + G model or the optimal partitioned model suggested by PartitionFinder. The optimal partitioned model used 29 gene-specific models (28 LG + G and one WAG + G). The AIC was used to test whether using a single model or a partitioned model provided a better fit to the data. Rate variation across lineages was modelled using both an autocorrelated and uncorrelated clock model. Bayesian cross-validation was used to test whether one of the two considered, relaxed molecular clock models best fitted the data (implemented in PhyloBayes).

In all our molecular clock analyses, we applied a soft tail of 2.5% to the upper calibration bound and a hard minimum, apart from the root node (to which a hard maximum was applied) and the nodes calibrated using *Bangiomorpha*[72] (to which a soft minimum tail of 2.5% was applied). For all molecular clock analyses, convergence was tested in Tracer[73] by comparing plots of estimates from the two independent chains and evaluating whether—for each model parameter and divergence time estimate—the effective sample size was sufficiently large. All reported molecular clock analyses reached excellent levels of convergence.

Reporting Summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this article.

Data availability. All data that support the findings of this study are available from Bitbucket: https://bitbucket.org/bzxdp/betts_et_al_2017.

Received: 15 April 2018; Accepted: 13 July 2018;
Published online: 20 August 2018

References

1. Dos Reis, M., Donoghue, P. C. J. & Yang, Z. Bayesian molecular clock dating of species divergences in the genomics era. *Nat. Rev. Genet.* **17**, 71–80 (2016).
2. Wacey, D. *Early Life on Earth: a Practical Guide* Vol. 31 (Springer, New York, 2009).
3. Inoue, J., Donoghue, P. C. J. & Yang, Z. The impact of the representation of fossil calibrations on Bayesian estimation of species divergence times. *Syst. Biol.* **59**, 74–89 (2009).
4. Parham, J. F. et al. Best practices for justifying fossil calibrations. *Syst. Biol.* **61**, 346–359 (2012).
5. Warnock, R. C. M., Parham, J. F., Joyce, W. G., Lyson, T. R. & Donoghue, P. C. J. Calibration uncertainty in molecular dating analyses: there is no substitute for the prior evaluation of time priors. *Proc. R. Soc. B* **282**, 20141013 (2014).
6. Davín, A. A. et al. Gene transfers can date the tree of life. *Nat. Ecol. Evol.* **2**, 904–909 (2018).
7. Lozano-Fernandez, J., dos Reis, M., Donoghue, P. C. J. & Pisani, D. RelTime rates collapse to a strict clock when estimating the timeline of animal diversification. *Genome Biol. Evol.* **9**, 1320–1328 (2017).
8. Pisani, D. & Liu, A. G. Animal evolution: only rocks can set the clock. *Curr. Biol.* **25**, R1079–R1081 (2015).
9. Spang, A. et al. Complex archaea that bridge the gap between prokaryotes and eukaryotes. *Nature* **521**, 173–179 (2015).
10. Williams, T. A., Foster, P. G., Cox, C. J. & Embley, T. M. An archaeal origin of eukaryotes supports only two primary domains of life. *Nature* **504**, 231–236 (2013).
11. Dodd, M. S. et al. Evidence for early life in Earth's oldest hydrothermal vent precipitates. *Nature* **543**, 60–64 (2017).
12. Pflug, H. D. & Jaeschke-Boyer, H. Combined structural and chemical analysis of 3,800-Myr-old microfossils. *Nature* **280**, 483–486 (1979).
13. Nutman, A. P., Bennett, V. C., Friend, C. R. L., Van Kranendonk, M. J. & Chivas, A. R. Rapid emergence of life shown by discovery of 3,700-million-year-old microbial structures. *Nature* **537**, 535–538 (2016).
14. Rosing, M. T. ^{13}C-depleted carbon microparticles in >3700-Ma sea-floor sedimentary rocks from West Greenland. *Science* **283**, 674–676 (1999).
15. Mojzsis, S. J.et al. Evidence for life on Earth before 3,800 million years ago. *Nature* **384,** 55–59 1996).
16. Van Zuilen, M. A., Lepland, A. & Arrhenius, G. Reassessing the evidence for the earliest traces of life. *Nature* **418**, 627–630 (2002).
17. Horita, J. & Berndt, M. E. Abiogenic methane formation and isotopic fractionation under hydrothermal conditions. *Science* **285**, 1055–1057 (1999).
18. Lepland, A., Arrhenius, G. & Cornell, D. Apatite in early Archean Isua supracrustal rocks, southern West Greenland: its origin, association with graphite and potential as a biomarker. *Precambrian Res.* **118**, 221–241 (2002).
19. Sugitani, K. et al. Early evolution of large micro-organisms with cytological complexity revealed by microanalyses of 3.4 Ga organic-walled microfossils. *Geobiology* **13**, 507–521 (2015).
20. Sugitani, K. et al. Biogenicity of morphologically diverse carbonaceous microstructures from the ca. 3400 Ma Strelley Pool Formation, in the Pilbara Craton, Western Australia. *Astrobiology* **10**, 899–920 (2010).
21. Sugitani, K., Mimura, K., Nagaoka, T., Lepot, K. & Takeuchi, M. Microfossil assemblage from the 3400 Ma Strelley Pool Formation in the Pilbara Craton, Western Australia: results form a new locality. *Precambrian Res.* **226,** 59–74 2013).
22. Wacey, D., Kilburn, M. R., Saunders, M., Cliff, J. & Brasier, M. D. Microfossils of sulphur-metabolizing cells in 3.4-billion-year-old rocks of Western Australia. *Nat. Geosci.* **4**, 698–702 (2011).
23. Lepot, K. et al. Texture-specific isotopic compositions in 3.4 Gyr old organic matter support selective preservation in cell-like structures. *Geochim. Cosmochim. Acta* **112**, 66–86 (2013).
24. Wacey, D., McLoughlin, N., Whitehouse, M. J. & Kilburn, M. R. Two coexisting sulfur metabolisms in a ca. 3400 Ma sandstone. *Geology* **38**, 1115–1118 (2010).
25. Wacey, D. Stromatolites in the ~3400 Ma Strelley Pool Formation, Western Australia: examining biogenicity from the macro-to the nano-scale. *Astrobiology* **10**, 381–395 (2010).
26. Abramov, O. & Mojzsis, S. J. Microbial habitability of the Hadean Earth during the late heavy bombardment. *Nature* **459**, 419–422 (2009).
27. Butterfield, N. J. *Bangiomorpha pubescens* n. gen., n. sp.: implications for the evolution of sex, multicellularity, and the Mesoproterozoic/Neoproterozoic radiation of eukaryotes. *Paleobiology* **26**, 386–404 (2000).
28. Sánchez-Baracaldo, P., Raven, J. A., Pisani, D. & Knoll, A. H. Early photosynthetic eukaryotes inhabited low-salinity habitats. *Proc. Natl Acad. Sci. USA* **114**, E7737–E7745 (2017).
29. Bengston, S. et al. Three-dimensional preservation of cellular and subcellular structures suggests 1.6 billion-year-old crown-group red algae. *PLoS Biol.* **15**, e2000735 (2017).
30. Lamb, D. M., Awramik, S. M., Chapman, D. J. & Zhu, S. Evidence for eukaryotic diversification in the ~1800 million-year-old Changzhougou Formation, North China. *Precambrian Res.* **173**, 93–104 (2009).
31. Zaremba-Niedzwiedzka, K. et al. Asgard archaea illuminate the origin of eukaryotic cellular complexity. *Nature* **541**, 353–358 (2017).
32. Warnock, R. C. M., Yang, Z. & Donoghue, P. C. J. Exploring uncertainty in the calibration of the molecular clock. *Biol. Lett.* **8**, 156–159 (2012).
33. Lanfear, R., Calcott, B., Ho, S. Y. W. & Guindon, S. PartitionFinder: combined selection of partitioning schemes and substitution models for phylogenetic analyses. *Mol. Biol. Evol.* **29**, 1695–1701 (2012).
34. Woese, C. R. & Fox, G. E. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proc. Natl Acad. Sci. USA* **74**, 5088–5090 (1977).
35. Drummond, A. J., Ho, S. Y. W., Phillips, M. J. & Rambaut, A. Relaxed phylogenetics and dating with confidence. *PLoS Biol.* **4**, e88 (2006).
36. Chapman, C. R., Cohen, B. A. & Grinspoon, D. H. What are the real constraints on the existence and magnitude of the late heavy bombardment? *Icarus* **189**, 233–245 (2007).
37. Hayes, J. M. in *Early life on Earth* Vol. 84, 220–236 (Columbia University Press, New York, 1994).
38. Ueno, Y., Yamada, K., Yoshida, N., Maruyama, S. & Isozaki, Y. Evidence from fluid inclusions for microbial methanogenesis in the early Archaean era. *Nature* **440**, 516–519 (2006).
39. Wolfe, J. & Fournier, G. P. Horizontal gene transfer constrains the timing of methanogen evolution. *Nat. Ecol. Evol.* **2**, 897–903 (2018).
40. Evans, P. N. et al. Methane metabolism in the archaeal phylum Bathyarchaeota revealed by genome-centric metagenomics. *Science* **350**, 434–438 (2015).
41. Vanwonterghem, I. et al. Methylotrophic methanogenesis discovered in the archaeal phylum Verstraetearchaeota. *Nat. Microbiol.* **1**, 16170 (2016).
42. Williams, T. A. et al. Integrative modeling of gene and genome evolution roots the archaeal tree of life. *Proc. Natl Acad. Sci. USA* **114**, 4602–4611 (2017).
43. Weiss, M. C. et al. The physiology and habitat of the last universal common ancestor. *Nat. Microbiol.* **1**, 16116 (2016).
44. Sousa, F. L., Nelson-Sathi, S. & Martin, W. F. One step beyond a ribosome: the ancient anaerobic core. *Biochim. Biophys. Acta* **1857**, 1027–1038 (2016).
45. Borrel, G., Adam, P. S. & Gribaldo, S. Methanogenesis and the Wood–Ljungdahl pathway: an ancient, versatile, and fragile association. *Genome Biol. Evol.* **8**, 1706–1711 (2016).
46. Lyons, T. W., Reinhard, C. T. & Planavsky, N. J. The rise of oxygen in Earth's early ocean and atmosphere. *Nature* **506**, 307–315 (2014).
47. Schirrmeister, B. E., de Vos, J. M., Antonelli, A. & Bagheri, H. C. Evolution of multicellularity coincided with increased diversification of cyanobacteria and the Great Oxidation Event. *Proc. Natl Acad. Sci. USA* **110**, 1791–1796 (2013).
48. Knoll, A. H. & Nowak, M. A. The timetable of evolution. *Sci. Adv.* **3**, e1603076 (2017).
49. Shih, P. M., Hemp, J., Ward, L. M., Matzke, N. J. & Fischer, W. W. Crown group Oxyphotobacteria postdate the rise of oxygen. *Geobiology* **15**, 19–29 (2017).
50. Kurland, C. G., Collins, L. J. & Penny, D. Genomics and the irreducible nature of eukaryote cells. *Science* **312**, 1011–1014 (2006).
51. Chernikova, D., Motamedi, S., Csürös, M., Koonin, E. V. & Rogozin, I. B. A late origin of the extant eukaryotic diversity: divergence time estimates using rare genomic changes. *Biol. Direct* **6**, 26 (2011).
52. Eme, L., Sharpe, S. C., Brown, M. W. & Roger, A. J. On the age of eukaryotes: evaluating evidence from fossils and molecular clocks. *CSH Perspect. Biol.* **6**, a016139 (2014).
53. Parfrey, L. W., Lahr, D. J. G., Knoll, A. H. & Katz, L. A. Estimating the timing of early eukaryotic diversification with multigene molecular clocks. *Proc. Natl Acad. Sci. USA* **108**, 13624–13629 (2011).
54. Shih, P. M. & Matzke, N. J. Primary endosymbiosis events date to the later Proterozoic with cross-calibrated phylogenetic dating of duplicated ATPase proteins. *Proc. Natl Acad. Sci. USA* **110**, 12355–12360 (2013).
55. McInerney, J. O., O'Connell, M. J. & Pisani, D. The hybrid nature of the Eukaryota and a consilient view of life on Earth. *Nat. Rev. Microbiol.* **12**, 449–455 (2014).
56. Roger, A. J., Muñoz-Gómez, S. A. & Kamikawa, R. The origin and diversification of mitochondria. *Curr. Biol.* **27**, R1177–R1192 (2017).
57. Pittis, A. A. & Gabaldón, T. Late acquisition of mitochondria by a host with chimeric prokaryotic ancestry. *Nature* **531**, 101–104 (2016).
58. Martin, W. F. et al. Late mitochondrial origin is an artifact. *Genome Biol. Evol.* **9**, 373–379 (2017).
59. Pisani, D., Cotton, J. A. & McInerney, J. O. Supertrees disentangle the chimerical origin of eukaryotic genomes. *Mol. Biol. Evol.* **24**, 1752–1760 (2007).
60. Ku, C. et al. Endosymbiotic origin and differential loss of eukaryotic genes. *Nature* **524**, 427–432 (2015).
61. Lane, N. & Martin, W. The energetics of genome complexity. *Nature* **467**, 929–934 (2010).

62. Benson, D. A. et al. GenBank. *Nucleic Acids Res.* **41**, D36–D42 (2013).
63. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
64. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).
65. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).
66. Lartillot, N., Lepage, T. & Blanquart, S. PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. *Bioinformatics* **25**, 2286–2288 (2009).
67. Kück, P. & Meusemann, K. FASconCAT: convenient handling of data matrices. *Mol. Phylogenet. Evol.* **56**, 1115–1118 (2010).
68. Lartillot, N. & Philippe, H. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Mol. Biol. Evol.* **21**, 1095–1109 (2004).
69. Aberer, A. J., Krompass, D. & Stamatakis, A. Pruning rogue taxa improves phylogenetic accuracy: an efficient algorithm and webservice. *Syst. Biol.* **62**, 162–166 (2013).
70. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007).
71. R Core Development Team. *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, 2017).
72. Butterfield, N. J., Knoll, A. H. & Swett, K. A bangiophyte red alga from the Proterozoic of arctic Canada. *Science* **250**, 104–108 (1990).
73. Rambaut, A., Drummond, A. J., Xie, D., Baele, G. & Suchard, M. A. Posterior summarisation in Bayesian phylogenetics using Tracer 1.7. *Syst. Biol.* https://doi.org/10.1093/sysbio/syy032 (2018).
74. Martijn, J., Vosseberg, J., Guy, L., Offre, P. & Ettema, T. J. G. Deep mitochondrial origin outside the sampled alphaproteobacteria. *Nature* **557**, 101–105 (2018).
75. Esser, C. et al. A genome phylogeny for mitochondria among alpha-proteobacteria and a predominantly eubacterial ancestry of yeast nuclear genes. *Mol. Biol. Evol.* **21**, 1643–1660 (2004).
76. Fitzpatrick, D. A., Creevey, C. J. & McInerney, J. O. Genome phylogenies indicate a meaningful α-proteobacterial phylogeny and support a grouping of the mitochondria with the Rickettsiales. *Mol. Biol. Evol.* **23**, 74–85 (2006).

Acknowledgements

H.C.B. was supported by a NERC GW4 PhD studentship. J.W.C. was supported by a BBSRC SWBio PhD studentship. M.N.P. was supported by an 1851 Royal Commission Fellowship. P.C.J.D. was supported by BBSRC grant BB/N000919/1. T.A.W. is supported by a Royal Society Fellowship and NERC grant NE/P00251X/1.

Author contributions

D.P., P.C.J.D. and T.A.W. designed the study. H.C.B. assembled the datasets and performed the phylogenetic and molecular clock analyses. M.N.P. and J.W.C. contributed further molecular clock analyses. H.C.B., D.P., P.C.J.D. and T.A.W. wrote the manuscript. All authors edited the manuscript and approved the final version.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41559-018-0644-x.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to D.P.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

nature ecology & evolution

SUPPLEMENTARY INFORMATION

https://doi.org/10.1038/s41559-018-0644-x

In the format provided by the authors and unedited.

Integrated genomic and fossil evidence illuminates life's early evolution and eukaryote origin

Holly C. Betts[1], Mark N. Puttick[1,2], James W. Clark[1], Tom A. Williams[1,3], Philip C. J. Donoghue[1] and Davide Pisani[1,3]*

[1]School of Earth Sciences, University of Bristol, Bristol, UK. [2]Milner Centre for Evolution, Department of Biology and Biochemistry, University of Bath, Bath, UK. [3]School of Biological Sciences, University of Bristol, Bristol, UK. *e-mail: davide.pisani@bristol.ac.uk

Supplementary information

S1. Calibration information

Node: Last universal common ancestor (LUCA)

Locality and Stratigraphy level: Strelley Pool Formation, Western Australia

Minimum age: 3347 Ma (3350 Ma ± 3 Myr[1])

Maximum age: 4520 Ma (4510 Ma ± 10 Myr[2,3])

Phylogenetic justification:

There are numerous reports of fossils from early Archaean sediments, however, determining a biotic origin for these records is difficult. Generally, there is a dearth of strata representative of early Earth history; those strata that are representative and are available for sampling have often been heavily altered by metamorphic processes. The oldest rocks available include, the Itsaq Gneiss, Isua, Greenland; the Barberton Greenstone Belt, South Africa; and the Pilbara Craton, Australia. These contain the oldest possible remains of life. At >3.7 Ga the Itsaq Gneiss contains putative fossils[4,5], stromatolites[6], carbon isotopes[7] and graphite inclusions[8,9]. However, each of these records has been disputed, either considered unlikely to be fossils, or that the record could be produced by geological rather than biological means[10-12] i.e. isotope ratios and graphite inclusions, synthesized by Fisher-Tropsch type (FTT) reactions[13,14].

At Pilbara, there are claims of isotopic evidence for sulphur bacteria[15], putative stromatolites and the infamous microfossils from the Apex Chert[16], as well as other microfossil reports[17,18]. None of these records is conclusive, when re-examined the Apex Chert microfossils[16] proved more likely to be an artefact of the reorganization of carbonaceous matter[19,20]. Likewise, the other microfossils have not been rigorously examined and so do not provide conclusive evidence of life. The sulphur isotope data[15] is also uncertain as it is possible to produce the

same signals by non-biological means[21]. Microfossils have also been reported from Barberton[22-25] but their biogenesis has been disputed.

Putative stromatolites are widespread in ancient sediments in both Barberton and Pilbara[26-31] but their formation is not exclusively tied to the presence of biological processes and the oldest stromatolites are most often found without any accompanying microbial fossils. Their abiogenic synthesis has been replicated laboratory conditions[33] and so they provide an uncertain record. Therefore, we must look for more conclusive evidence of life, that which has been examined from several angles. More rigorous analysis has been undertaken of fossils from slightly younger sites. For example, a sample of fossils from the ~3.2 Ga Moodies Group, Barberton, were described using criteria which looked at a rigorous range of criteria: fossil placement within the rock; their ultrastructure; their composition; and their size[34]. Some of these small organic walled fossils are actually very large (up to 300 microns diameter)[34]; sizes which are unknown amongst archaea[35]. Older remains from the Strelley Pool Formation, Pilbara, Western Australia[36,37] have also been examined based on a set of criteria similar to those used by Javaux and colleagues. These fossils have a complex ultrastructure and acid resistant walls that survive being digested out of the rock. Additionally, it should be noted that the organic carbon signature shows that the fossils were not emplaced into the rock at a later stage, a problem with many early records. Some of these fossils are also present in multi-cell chains. These are not known to form in abiotic ways and, hence, it can be concluded that these structures are biological in origin. The Strelley Pool Formation also contains a host of other evidence for life. These include other microfossils both alone[38] and in association with pyrite crystals[39], possibly indicating some kind of sulphur metabolism backed up a previous study showing sulphur metabolism[40], as well as microbial mats[41], and stromatolites, which have been more intensely studied to give credence to their biological affinity[42]. What is more the microfossils have been shown to possess specific $\delta^{13}C_{org}$ signatures that are correlated

specifically to the microfossils[43]. Overall these show a diverse community[44]. Although alone these would not provide a suitable record, in accordance with the well-studied fossils[36] they provide a robust calibration with which to constrain LUCA.

Age justification:

Hard minimum: The Strelley Pool Formation is located in North Eastern Australia and is part of the larger Pilbara Craton. The stratigraphic position of this formation (also known as the Strelley Pool Chert) has been contentious but it is now argued to form a layer between the Warawoona and Kelly groups[45]. The formation is dated to 3.426-3.350 Ga[45], with the minimum age (3.350 Ga ± 0.003 Gyr) based on a volcaniclastic tuff, at the base of the overlying Euro Basalt[1] in the Kelly Group. Hence our minimum age constraint is 3.347 Ga.

Soft maximum: We can use the Moon-forming impact as a maximum constraint; there is no other event or date of significance which can be used in its place. This devastating event would have sterilised the Earth, hence any life now present on the planet must have evolved post-impact. It has been proposed that life would not have been able to survive the late heavy bombardment, which post-dated the Moon-forming impact, but this view has been contested as ideas of a cool early earth and an early ocean have been proposed[46,47], as well as models which show that life would have been able to survive during this intense bombardment[48]. It is also possible that there was no late heavy bombardment because evidence of its occurrence has been found on the Moon but not on Earth[49]. There is some debate over the exact timing of the impact with proposed dates ranging from 4.54 Ga ± 0.01 Myr[50] to ~4.44Ga[51]. Some of the most recent simulations and models place the Moon-forming impact at ~4.47 Ga based on asteroidal meteorites and siderophile elements[52,53]. This concurs with estimates based on U-Pb isotopes[54], Hf/W isotopes[55] and Rb/Sr isotopes[56]. We use the oldest credible date to encompass reasonable uncertainty. The oldest date of 5.4 Ga is based on the Hf-W system[50,57], around which there is some debate as to the amount of signal caused by cosmogenic production of ^{182}W from ^{181}Ta[58].

Hence, the most credible date comes from the U-Pb system. We follow other critical reviewers[59] in accepting Pb-Pb dating carried out on Moon rocks, yielding a date of 4.51 Ga ± 10 Myr [2]: a date which has also recently been confirmed by reanalysis of the Apollo zircons[3]. Thus, our maximum constraint is 4.52 Ga.

Node: Total group Cyanobacteria

Locality and Stratigraphy level: Manzimnyama Banded Ironstone Formation, Fig Tree Group, Barberton, South Africa

Minimum age: 3225 Ma (3226 Ma ± 1 Myr[60])

Maximum age: 4520 Ma (4510 Ma ± 10 Myr[2])

Phylogenetic justification: Cyanobacteria are the only living group of organisms that have evolved oxygenic photosynthesis. Proposed records of cyanobacteria from ancient rocks include Banded Ironstone Formations (BIFs), stromatolites, biomarkers, and a number of isotope systems. BIFs, which are found among the oldest sedimentary rocks, including protoliths of the 3.8 Ga Itsaq Gneiss, show a reduction of ferrous iron which has been claimed to occur due to cyanobacterial effects. However, arguments have been presented for the production of BIFs via abiogenic ultra-violet induced photolysis[61] and anoxygenic bacterial photosynthesis[62,63]. Early stromatolites are not sufficient evidence as they are not all biogenic and they don't necessarily require cyanobacteria for formation[32,64]. The best indicator of free oxygen at levels incompatible with photolysis, is from isotopes. These are a good proxy for oxygen because many elements are very sensitive to oxidative weathering. Prior to the Great Oxygenation Event, oxygen records in the form of isotopes extend back to 3.25 Gyr[65]. The authors report stable Fe and U-Th-Pb isotopes from the Manzimnyama BIF in the Fig Tree Group, Barberton, South Africa, which indicate a level of free oxygen indicative of cyanobacterial activity. They also find that there is a stratification in oxygen levels at the site,

showing an oxygenated shallow water layer and an anoxic deeper water. They argue that this is what we would expect to see in areas where there is some cyanobacterial activity. It is possible that oxygen was being produced in smaller quantities prior to the GOE and that these pockets of oxygen could be concentrated in an otherwise anoxic water column[66]. Other evidence for oxygenation from within this sequence comes from the Moodies group which lies immediately above the Fig Tree Group at Barberton. This has macroscopic tufted microbial mats[67], that are thought to grow upwards towards a source of light, and in modern examples are made mostly of cyanobacteria. Additionally, this evidence for oxygenation is not isolated as numerous other lines of evidence, based mainly upon redox sensitive elements and other isotopes, now support the appearance of pre-GOE oxygen being produced by cyanobacteria[68-73].

Age justification:

Hard minimum: The isotopic evidence from the Manzimnyama BIF in the Fig Tree Group, Barberton, South Africa[65]. The age of the Fig Tree Group is well constrained with a spherule layer at its base dated at 3258 Ma ± 3 Myr[74], and an overlying volcanic unit at its top dated at 3226 Ma ± 1 Myr[60]. Hence, the minimum date we would assign is 3225 Myr.

Soft maximum: We can use the Moon-forming impact as a maximum constraint, as there no other event or date of significance which can be used in its place. This devastating event would have sterilised the Earth, hence any life now present on the planet must have evolved post-impact. It has been proposed that life would not have been able to survive the late heavy bombardment, which post-dated the Moon-forming impact, but this view has been contested as ideas of a cool early earth and an early ocean have been proposed[46,47] as well as models which show that life would have been able to survive during this intense bombardment[48]. It is also possible that there was no late heavy bombardment because evidence of its occurrence has been found on the Moon but not on Earth[49]. There is some debate over the exact timing of the

126 impact with proposed dates ranging from 4.54 Ga ± 0.01 Myr[50] to ~4.44Ga[51]. Some of the most
127 recent simulations and models place the Moon-forming at ~4.47 Ga based on asteroidal
128 meteorites and siderophile elements[52,53]. This concurs with estimates based on U-Pb isotopes[54],
129 Hf/W isotopes[55] and Rb/Sr isotopes[56]. We use the oldest credible date to encompass reasonable
130 uncertainty. The oldest date of 5.4 Ga is based on the Hf-W system[50,57], around which there is
131 some debate as to the amount of signal caused by cosmogenic production of ^{182}W from ^{181}Ta[58].
132 Hence, the most credible date comes from the U-Pb system. We follow other critical
133 reviewers[59], in accepting Pb-Pb dating carried out on Moon rocks yields a date of 4.51 Ga ± 10
134 Myr [2] a date which has also recently been confirmed by reanalysis of the Apollo zircons[3]. Thus,
135 our maximum constraint is 4.52 Ga.
136
137 **Node:** Total group Eukarya
138 **Locality and Stratigraphy level:** Changcheng Group, Hebei Province, North China
139 **Minimum age:** 1619.1 Ma (1625.3 ± 6.2 Myr[75])
140 **Maximum age:** 4520 Ma (4510 Ma ± 10 Myr[2])
141 **Phylogenetic justification:**
142 The record of eukaryotes covers a large timespan, during much of which the fossils attributed
143 to eukaryotes are relatively simple and do not exhibit much morphological variation. The
144 earliest of these that have been rigorously examined are those from the Changcheng Group in
145 North China. These fossils come from two levels within this group, the Changzhougou Fm.
146 and the Chuanlinggou Fm[76,77]. The units are made up of sandstone and shale, within which the
147 fossils are found. The fossils are small and lenticular in shape with a carbonaceous outer sheath
148 and what are interpreted to be excystment structures. The complexity exhibited by these sheaths
149 and the inferred function, along with the size, places them into the eukaryote domain. The
150 forms preserved at Changcheng are large enough, on average >125μm that they unlikely to be

any kind of Euacteria or Archaeabacter. Some bacterial cells can reach large sizes and size is not the best criteria to use but can be informative when used in conjunction with other characteristics. The authors demonstrate that the cells have a double sheath. The possibility that cyanobacteria have these structures is discussed but refuted on the basis of size. They are even proposed to be part of the green-algae plant lineage[78]. However, it is due to a lack of definitive features this claim cannot be substantiated. The age of these fossils encompasses reports of other fossils that are also Eukaryotic in nature, but those which also have uncertain affinities, such as the probable 1.56 Ga multicellular fossils[79], the string of beads *Horodyskia*[80], and *Shuiyousphaeridium*[81] and other acritarch and leiosphaerid forms[82,83]. Unfortunately, these fossils are not diagnostic of any crown group eukaryotes and so we can only use them to calibrate the total group of eukaryotes, helping us to provide a robust minimum for their appearance. Putative rhodophytes from the Chitrakoot Formation are slightly younger (see total-group Rhodophyta, below). Although some are sceptical of the eukaryotic nature of these fossils[84], the combination of their morphology and size seems sufficient to assign them to a stem group eukaryote affinity.

Age justification:

Hard minimum:

As the oldest of these fossils are found in the Changzhougou Formation it is this that we can date. To acquire a minimum date for the whole formation, we use ash layers in the overlying formation, yielding an age of 1625.3 ± 6.2 Myr[75]. The microfossils are present in both these layers, but have been described separately[76,77]. Hence, we can use the date of the oldest Chuanlinggou, 1619.1 Ma, to date the underlying Changzhougou.

Soft maximum: We can use the Moon-forming impact as a maximum constraint, as there no other event or date of significance which can be used in its place. This devastating event would have sterilised the Earth, hence any life now present on the planet must have evolved post-

impact. It has been proposed that life would not have been able to survive the late heavy bombardment, which post-dated the Moon-forming impact, but this view has been contested as ideas of a cool early earth and an early ocean have been proposed[46,47] as well as models which show that life would have been able to survive during this intense bombardment[48]. It is also possible that there was no late heavy bombardment because evidence of its occurrence has been found on the Moon but not on Earth[49]. There is some debate over the exact timing of the impact with proposed dates ranging from 4.54 Ga ± 0.01 Myr[50] to ~4.44Ga[51]. Some of the most recent simulations and models place the Moon-formation at ~4.47 Ga based on asteroidal meteorites and siderophile elements[52,53]. This concurs with estimates based on U-Pb isotopes[54], Hf/W isotopes[55] and Rb/Sr isotopes[56]. We use the oldest credible date to encompass reasonable uncertainty. The oldest date of 5.4 Ga is based on the Hf-W system[50,57], around which there is some debate as to the amount of signal caused by cosmogenic production of ^{182}W from ^{181}Ta[58]. Hence, the most credible date comes from the U-Pb system. We follow other critical reviewers[59], in accepting Pb-Pb dating carried out on Moon rocks yields a date of 4.51 Ga ± 10 Myr [2] a date which has also recently been confirmed by reanalysis of the Apollo zircons[3]. Thus, our maximum constraint is 4.52 Ga.

Node: Total group Rhodophyta

Specimen and fossil taxon: *Bangiomorpha pubescens.* (Holotype) HUPC 62912, Slide HUST-1A, England Finder coordinates: O-35.

Locality and Stratigraphy level: Lower Hunting Formation, Somerset Island, arctic Canada.

Soft Minimum age: 1033 Ma (1092 Ma ± 59 Myr[86])

Soft Maximum age: 1891 Ma (1823 Ma ± 68 Myr[85])

Phylogenetic justification: There are several reports of red algae within the fossil record, stretching back into the Ediacaran, Neo- and Meso-proterozoic. The oldest of which are 1.6

billion year old fossils, *Rafatazmia chitrakootia* and *Ramathallus lobatus*, from the Chitrakoot Formation[87]. However, though both are suggested to be red algae and, while the remains are compatible with this interpretation, they no not preclude alternative assignments within total group Archaeplastida. *Bangiomorpha pubescens* is younger fossil, originally described as a Bangiale red algae in comparison to the extant *Bangia* due to the distinctive, radially orientated, intercalary division of its cells and its putative development[88,89]. It has therefore been used as a calibration for the red algae or sometimes more specifically for the bangiophyte red algae[90,91]. Red algae are united by general characteristics that are not commonly preserved in the fossil record, even in the most exceptional of circumstances, e.g. the red coloured pigments, and unstacked thylakoids within the chloroplasts[92,93]. Hence, *Bangiomorpha* was identified using potential developmental characters and the distinct shape of its cell arrangements. However, although these characters are distinctive[92], they are also characteristic of several other red algae[94]. *Bangiomorpha* has been described as having a multicellular holdfast, a feature found in some Compsopogonophyceae, another group of basal red algae. Modern *Bangia* has an attachment rhizoid, not a multicellular holdfast indicating that the features of *Bangiomorpha* are not specifically Bangiale. These observations make it inappropriate to assign *Bangiomorpha* specifically to Bangiales. However, the distinct developmental, reproductive and morphological characteristics appear sufficient to assign *Bangiomorpha* to Rhodophyta as a whole. Hence, we can use this fossil to calibrate the node subtending Rhodophyta which link them to their nearest common ancestor.

Age justification:

Soft minimum constraint: The oldest records of *Bangiomorpha pubescens* occur in the Lower Hunting Formation, of Somerset Island, Arctic Canada. A minimum age for the formation is based on the age of the Franklin igneous events, which have been dated to 723 Ma ± 3 Myr[95], with a maximum age of 1267 Ma ± 2 Myr based on the McKenzie igneous events[96]. The

original description[89] cites an unpublished Pb-Pb date 1198 Ma ± 24 Myr as a best date for *B. pubescens*, however, this date remains unsubstantiated and so it must be discounted. The formation from which *Bangiomorpha* was recovered can be correlated lithostratigraphically to the Society Cliffs Formation[97] and the Uluskan Group[98], which are closer to the base of the sequence, and dated at ~1267 Ma (Mesoproterozoic). This is substantially older than the ~723 Ma minimum constraint on the age of the Lower Hunting Formation. The other option is a date of 1092 ± 59 Myr[86] established from a shale layer present in the Arctic Bay formation, which is comparable[99] to the sequences below the *Bangiomorpha* fossiliferous layer i.e. the Lower Hunting formation. Although this date is older than the layer in which *Bangiomorpha* resides it is very close in age and so we employ it as a soft-minimum constraint, thus our date for this fossil is 1033 Ma.

Soft Maximum Constraint: The soft maximum constraint is based on the earliest record of eukaryotes[76,77,100] when, despite the presence of simple eukaryotes, there is no evidence of anything as complex as a definitively multicellular alga. Though the fossils present have been suggested by some to represent some kind of green algae[78]. The maximum for this formation is based on the igneous and metamorphic rocks that it overlies. These rocks are dated at 1823 Ma ± 68 Myr[85], yielding a soft maximum constraint of 1891 Ma.

Nodes: Crown Alphaproteobacteria

Specimen and fossil taxon: *Bangiomorpha pubescens.* (Holotype) HUPC 62912, Slide HUST-1A, England Finder coordinates: O-35.

Locality and Stratigraphy level: Lower Hunting Formation, Somerset Island, arctic Canada.

Soft Minimum age: 1033 Ma (1092 Ma ± 59 Myr[86])

Soft Maximum age: 4520 Ma (4510 Ma ± 10 Myr[2])

Phylogenetic justification: There are no fossils that can be attributed to Alphaproteobacteria. However, the important eukaryote organelle, the mitochondria has been found by consensus to have belonged within Alphaproteobacteria. This is because mitochondria formed via an endosymbiosis event with the protoeukaryote[101]. Within the alphaproteobacteria group the mitochondria are most commonly linked to the *Rickettsiales*[102,103] though arguments have also been made for them belonging to other alphaproteobacterial groups[101,104,105]. Mitochondria contain a mosaic of genes which are not all alphaproteobacterial in origin[106,107], but nonetheless it is still believed to have originated within this group. *Bangiomorpha pubescens*[88] is a total group rhodophyte with features that link it to the basal rhodophyte groups such as its cell arrangement, and others which mean it cannot be placed specifically within any one of them. It is the oldest fossil in the record that can be confidently identified as a crown-eukaryote. There are older fossils that are eukaryotic in nature, but they cannot be placed with certainty into crown-Eukaryota. Hence, we can use Bangiomorpha to provide some level of constraint to the alphaproteobacteria, in a part of the tree of life that is otherwise poorly constrained.

Age justification:

Soft minimum constraint: The oldest records of *Bangiomorpha pubescens* occur in the Lower Hunting Formation, of Somerset Island, Arctic Canada. A minimum age for the formation is based on the age of the Franklin igneous events, which have been dated to 723 Ma ± 3 Myr[95], with a maximum age of 1267 Ma ± 2 Myr based on the McKenzie igneous events[96]. The original description[89] cites an unpublished Pb-Pb date 1198 Ma ± 24 Myr as a best date for *B. pubescens*, however, this date remains unsubstantiated and so it must be discounted. The formation from which *Bangiomorpha* was recovered can be correlated lithostratigraphically to the Society Cliffs Formation[97] and the Uluskan Group[98], which are closer to the base of the sequence, and dated at ~1267 Ma (Mesoproterozoic). This is substantially older than the ~723 Ma minimum constraint on the age of the Lower Hunting Formation. The other option is a date

of 1092 ± 59 Myr[86] established from a shale layer present in the Arctic Bay formation, which is comparable[99] to the sequences below the *Bangiomorpha* fossiliferous layer i.e. the Lower Hunting formation. Although this date is older than the layer in which *Bangiomorpha* resides it is very close in age and so we employ it as a soft-minimum constraint, thus, our minimum for this clade is 1033 Ma.

Soft maximum: We can use the Moon-forming impact as a maximum constraint, as there no other event or date of significance which can be used in its place. This devastating event would have sterilised the Earth, hence any life now present on the planet must have evolved post-impact. It has been proposed that life would not have been able to survive the late heavy bombardment, which post-dated the Moon-forming impact, but this view has been contested as ideas of a cool early Earth and an early ocean have been proposed[46,47] as well as models which show that life would have been able to survive during this intense bombardment[48]. It is also possible that there was no late heavy bombardment because evidence of its occurrence has been found on the Moon but not on Earth[49]. There is some debate over the exact timing of the impact with proposed dates ranging from 4.54 Ga ± 0.01 Myr[50] to ~4.44Ga[51]. Some of the most recent simulations and models place the Moon formation at ~4.47 Ga based on asteroidal meteorites and siderophile elements[52,53]. This concurs with estimates based on U-Pb isotopes[54], Hf/W isotopes[55] and Rb/Sr isotopes[56]. We use the oldest credible date to encompass reasonable uncertainty. The oldest date of 5.4 Ga is based on the Hf-W system[50,57], around which there is some debate as to the amount of signal caused by cosmogenic production of ^{182}W from ^{181}Ta[58]. Hence, the most credible date comes from the U-Pb system. We follow other critical reviewers[59], in accepting Pb-Pb dating carried out on Moon rocks yields a date of 4.51 Ga ± 10 Myr [2] a date which has also recently been confirmed by reanalysis of the Apollo zircons[3]. Thus, our maximum constraint on the age of Alphaproteobacteria is 4.52 Ga.

Nodes: Crown-Cyanobacteria

Specimen and fossil taxon: *Bangiomorpha pubescens.* (Holotype) HUPC 62912, Slide HUST-1A, England Finder coordinates: O-35.

Locality and Stratigraphy level: Lower Hunting Formation, Somerset Island, arctic Canada.

Soft Minimum age: 1033 Ma (1092 Ma ± 59 Myr[86])

Soft Maximum age: 4520 Ma (4510 Ma ± 10 Myr[2])

Phylogenetic justification: Cyanobacteria are inferred to have a relatively plentiful fossil record. Often the Great Oxidation Event (GOE) and a number of fossils are used to calibrate the origins of the crown group and various lineages within it. However, the assumption that the GOE was caused by crown cyanobacteria rests on the assumption that photosynthesis evolved in associated with the crown clade. This has been recently challenged and so we do not use it as a calibration here[108]. Potential records of cyanobacteria extend into the Archaean but these are mainly simple cells and filaments[109] whose affinities cannot be substantiated[110]. There are fossils described as akinetes, cyanobacterial resting spores, from 21. Ga[111] and 1.6 Ga[112]. However, modern specimens show a range of characters and morphology making it difficult to relate these to any potential ancient counterparts, and other bacterial cells can also show this type of simple morphology[113]. The most convincing fossil remains are found in the Belcher Formation, Canada[114,115], from around 1.9 billion years old, however, even these cannot be discriminated confidently from other bacterial grades[113]. Instead of using the above-mentioned fossils as calibration points, as in other studies[116], we opted for a more conservative approached and used evidence for the oldest archaeplastid; this would have had a chloroplast, known to have originated in an endosymbiotic event with a cyanobacteria. There is no strict consensus as to which cyanobacterial group plastids evolved from with the main argument being whether they evolved from an early[117] or late[118,119] branching lineage within Cyanobacteria. *Bangiomorpha pubescens*[88] is a total group Rhodophyte (see total-group Rhodophyta, above).

It is the oldest fossil in the record that can be confidently identified as a crown group eukaryote; there are older fossils that are eukaryotic in nature, but they cannot be placed with any certainty into one of the extant eukaryotic groupings.

Age justification:

Soft minimum constraint: The oldest records of *Bangiomorpha pubescens* occur in the Lower Hunting Formation, of Somerset Island, Arctic Canada. A minimum age for the formation is based on the age of the Franklin igneous events, which have been dated to 723 Ma ± 3 Myr[95], with an maximum age of 1267 Ma ± 2 Myr based on the McKenzie igneous events[96]. The original description[89] cites an unpublished Pb-Pb date 1198 Ma ± 24 Myr as a best date for *B. pubescens*, however, this date remains unsubstantiated and so it must be discounted. The formation from which *Bangiomorpha* was recovered can be correlated lithostratigraphically to the Society Cliffs Formation[97] and the Uluskan Group[98], which are closer to the base of the sequence, and dated at ~1267 Ma (Mesoproterozoic). This is substantially older than the ~723 Ma minimum constraint on the age of the Lower Hunting Formation. The other option is a date of 1092 ± 59 Myr[86] established from a shale layer present in the Arctic Bay formation, which is comparable[99] to the sequences below the *Bangiomorpha* fossiliferous layer i.e. the Lower Hunting formation. Although this date is older than the layer in which *Bangiomorpha* resides it is very close in age and so we employ it as a soft-minimum constraint, thus, our minimum for this clade is 1033 Ma.

Soft maximum: We can use the Moon-forming impact as a maximum constraint, as there no other event or date of significance which can be used in its place. This devastating event would have sterilised the Earth, hence any life now present on the planet must have evolved post-impact. It has been proposed that life would not have been able to survive the late heavy bombardment, which post-dated the Moon-forming impact, but this view has been contested as ideas of a cool early earth and an early ocean have been proposed[46,47] as well as models

which show that life would have been able to survive during this intense bombardment[48]. It is also possible that there was no late heavy bombardment because evidence of its occurrence has been found on the Moon but not on Earth[49]. There is some debate over the exact timing of the impact with proposed dates ranging from 4.54 Ga ± 0.01 Myr[50] to ~4.44Ga[51]. Some of the most recent simulations and models place the Moon-forming at ~4.47 Ga based on asteroidal meteorites and siderophile elements[52,53]. This concurs with estimates based on U-Pb isotopes[54], Hf/W isotopes[55] and Rb/Sr isotopes[56]. We use the oldest credible date to encompass reasonable uncertainty. The oldest date of 5.4 Ga is based on the Hf-W system[50,57], around which there is some debate as to the amount of signal caused by cosmogenic production of ^{182}W from ^{181}Ta[58]. Hence, the most credible date comes from the U-Pb system. We follow other critical reviewers[59], in accepting Pb-Pb dating carried out on Moon rocks yields a date of 4.51 Ga ± 10 Myr [2] a date which has also recently been confirmed by reanalysis of the Apollo zircons[3]. Thus, our maximum constraint is 4.52 Ga.

Node: Dikarya

Locality and stratigraphy level: Rhynie, Aberdeenshire, Scotland. Lower Devonian

Minimum age: 392.1 Ma (393.3 Ma ± 1.2 Myr[120])

Maximum age: 1891 Ma (1823 Ma ± 68 Myr[85])

Phylogenetic justification: The minimum constraint is based upon fossils from the Rhynie Chert[121] described as *Paleopyrenomycites devonicus*[122]. This fungal fossil is found in association with the roots of early plants and has key characteristics that relate it to the Ascomycota, including containing the sexual spores (asci) in a sac-like structure, the ascus. Although there are earlier examples of possible fossil fungi much of their interpretation is spurious. This category includes *Tappania*, which was once interpreted as a fungus[123], but is now considered to be an acritarch[124], and the 'lichen-like' fossil from Doushantuo[125] is difficult

to discriminate from diagenetic artefacts that are characteristic of fossils from the Weng'an Biota[126]. There is a more convincing record of a possible Glomeromycota fungus from the Ordovician[127]. However, this specimen has not been assigned with as much confidence to a distinct fungal lineage as those fossils contained in the younger Devonian Rhynie Chert deposits. The oldest report of a fungi-like fossil is from the Ongeluk Formation, ~2.4 Ga[128]. The filaments are situated within basaltic lavas, a rock type shown to host putative fungal species in more recent Eocene basalts[129-131]. However, although the Ongeluk fossils do show many typical fungal features, these can also be attributed to the actinobacteria, such as the hyphae-like cells and Y-junctions, thus, their affinities are ambiguous. Hence, we use the confidently assigned fungi fossil from the Rhynie chert to constrain the minimum age of the clade comprising Ascomycota and Basidiomycota and sister lineage Glomeromycota.

Age justification:

Hard minimum: Proposed dates for the Rhynie Chert system have been mostly based upon zircons from volcanic deposits in the sequence. Two recent dates proposed are 407.1 Ma ± 2.2 Myr[132] and 411.5 Ma ± 1.3 Myr[133]. The former is from a hydrothermally produced layer within the sequence and with which there is high oxygen isotopic homogeneity from the layers with the spore bearing assemblage[132]. The other date is derived from the Milton of Noth andesite[133]. Despite being based on zircon evidence, neither of these dates is suitable; the Milton of Noth andesite has uncertain placement within the sequence but is most likely found beneath the Rhynie spore-bearing layer[134] and so cannot be used to provide a minimum date. The later date[132] is also unsuitable because the layers which are dated do not come from above the spore assemblage, and the method of dating has some problems[135]. Therefore, we base our minimum clade age constraint on the spore assemblage characterizing the Rhynie Chert. This places the Rhynie Chert in the early Pragian to early Emsian[136]. The age of the top of the Emsian-Eifelian

400 boundary is dated as 393.3 Ma ± 1.2 Myr[120]. Hence our minimum clade age constraint is 392.1
401 Ma.
402 **Soft maximum:** The maximum for this calibration is based on the earliest record of
403 eukaryotes[76,77,100] when, despite the presence of simple eukaryotes, there is no evidence of
404 anything as complex as a multicellular alga. Though the fossils present have been suggested
405 by some to represent some kind of green algae[78]. This date also encompasses the recent
406 discovery of possible multicellular eukaryotes from the 1.56 Ga[79]. The maximum for this
407 formation is based on the igneous and metamorphic rocks that lie beneath it. These rocks are
408 dated at 1823 Ma ± 68 Myr[85], thus, our maximum is 1891 Ma.
409
410 **Node:** Crown group Foraminifera
411 **Locality and Stratigraphy level:** The Chapel Island Formation, Newfoundland, Canada.
412 Lower Cambrian.
413 **Minimum age**: 525.5 Ma (525.5 Myr[120])
414 **Maximum age:** 1891 Ma (1823 Ma ± 68 Myr[85])
415 **Phylogenetic justification:**
416 The foraminifera are a group of testate eukaryotes that are part of Rhizaria, a group that also
417 includes Cercozoa and Radiolaria. Foraminifera are well known from most of the Proterozoic
418 before which there are scattered reports with varying degrees of validity. The very oldest
419 possible reports come from Post-Sturtian deposits located in Namibia and Mongolia[137,138].
420 These are interpreted as foraminifera based on the composition of the tests found. However,
421 the authors cautiously interpret them as foraminifera, partly due to the shape that is not seen in
422 modern forms, so there is still a level of uncertainty in their affinity. Other Ediacaran fossils
423 have been described as foraminifera, such as the enigmatic *Palaeopascichnus*. However, these
424 fossils lack a number of key diagnostic features of foraminifera[139]. Generally the oldest forms

425 are regarded to be those from Western African[140] and from the Lower Cambrian of Canada[141].
426 Though Culver described the Western African forms as Cambrian in nature, due to their
427 position and the appearance of a Cambrian snail in the same deposits, new dating suggests that
428 the formation might actually be closer to the Ordovician in age[142]. The fossil described as
429 *Platysolenites cooperi*[141] has had its foraminiferal affinity questioned based on the possible
430 composition of their tests[143,144]. However, in their paper McIlroy and colleagues dispel this
431 doubt by looking in detail at the wall composition. They find that it is composed of agglutinated
432 grains, was organically bound and probably flexible in life. They also find that it shows
433 evidence of fracturing that was repaired during the organism's lifetime, on the outside of the
434 wall, a character not seen in metazoans. This and other support from previous reviews[145-147]
435 provides strong evidence for *P. cooperi* being an early agglutinating foraminifera.
436 **Age justification:**
437 **Minimum:** The oldest fossils of *P. cooperi* come from the latest Ediacaran to Lower Cambrian
438 in Newfoundland, the Chapel Island formation[141]. This formation sits just above the Cambrian
439 boundary and is correlated to the Nemakit-Daldyian which has a minimum date of 525.5 Ma
440 according to the latest version of the geological timescale[120].
441 **Maximum:** The maximum for this calibration is based on the earliest record of
442 eukaryotes[76,77,100] recovered from the Changzhougou Formation (China), when, despite the
443 presence of simple eukaryotes, there is no evidence of crown-eukaryote lineages or their
444 characters. This date also encompasses the recent discovery of possible multicellular
445 eukaryotes from the 1.56 Ga[79] as well as the reports of possible ameboid tests, called vase-
446 shaped microfossils which might belong to a clade of the Rhizaria[137]. The maximum for the
447 Changcheng Group is based on the igneous and metamorphic rocks that lie beneath it. These
448 rocks are dated at 1823 ± 68 Myr[85], thus, our maximum is 1891 Ma.
449

Node: Embryophytes

Locality and Stratigraphy level: Qusaiba-1 core from the Quasim formation of northern Saudi Arabia

Minimum age: 448.5 Ma[149]

Maximum age: 509 Ma[149]

Age justification:

The oldest evidence of embryophytes are trilete spores. We follow Clark and Donoghue[149] in dating these to a minimum date of 448.5 Ma. The maximum is placed at the Bright Angel Shale which has a date of 507.2-509 Ma, hence, the maximum that we use to 509 Ma.

Node: Angiospermae

Locality and Stratigraphy level: Cowleaze Chine Member of the Vectis Formation of the Isle of Wight

Minimum age: 125.9 Ma (126.3 Ma ± 0.4 Myr[149])

Maximum age: 247.3 Ma (247.1 Ma ± 0.2 Myr[149])

Age justification:

The oldest evidence of angiosperms is tricopolate pollen. We follow Clark and Donoghue[149] and date the pollen to the Cowleaze Chine Member, Isle of White. This yields a minimum date of 126.3 ± 0.4 Myr and a maximum date of 247.1 Ma ± 0.2 Myr from a rock layer free of angiosperm pollen.

Node: Metazoa

Locality and Stratigraphy level: White Sea Formation, Russia

Minimum age: 550.25 Ma (552.85 Ma ± 2.6 Myr[150])

Maximum age: 833 Ma (827 Ma ± 6 Myr[150])

Age justification: The oldest uncontroversial evidence for Metazoa is the fossil *Kimberella quadrata*. The oldest specimen of this is found in the White Sea, Russia, for which a minimum date of 552.85 Ma ± 2.6 Myr has been established. The maximum is set as 827 Ma ± 6 from a formation that shows no evidence of any total group metazoans.

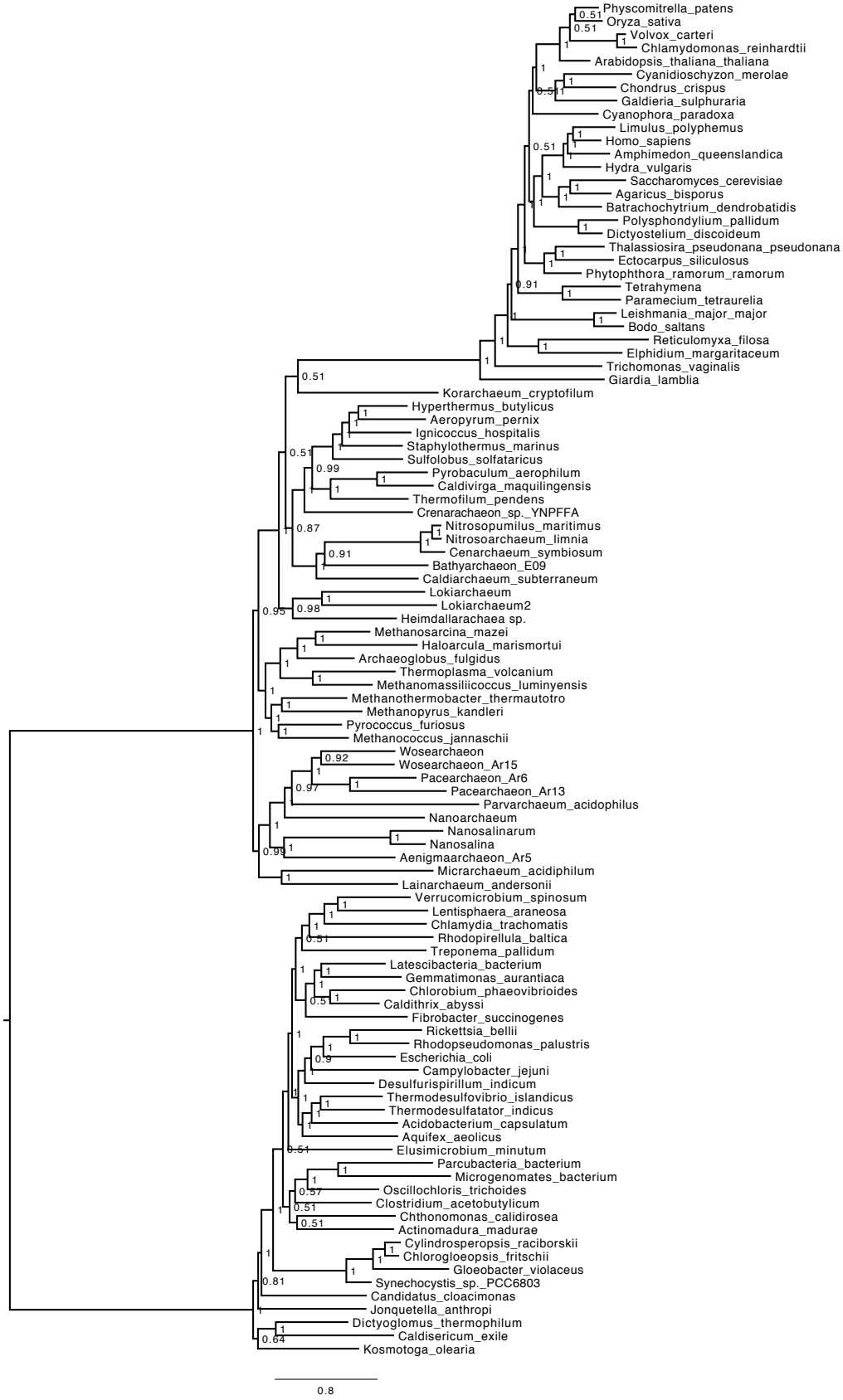
S2. Gene families used in this study by *S. cerevisiae* identification code.

S. cerevisiae gene IDs	Gene family number (arbitrary, corresponds to dm_XX.fa naming scheme)
Rps14bp	(1)
Rps23bp	(6)
Fun12p	(14)
Rpl11ap	(15)
Rsp3p	(20)
Rps16ap	(22)
Rpl1ap	(24)
Rpl2bp	(29)
Rpl23bp	(30)
Rpl12ap	(31)
Eft1p	(33)
Kae1p	(34)
Rps0bp	(35)
Rps2p	(36)
Rps5p	(37)
Srp54p	(40)
Tef1p	(4)
Rli1p	(5)
Dps1p	(10)
Rpa190p	(11)
Sec61p	(12)
Cct5p	(16)
Rfc2p	(17)
Vma2p	(23)
Map2p	(25)
Rpl16ap	(28)
Gln4p	(32)
Rpa135p	(39)
Srp101p	(41)

S3 Supplementary results – Phylogeny.

We performed phylogenetic analyses of our complete dataset to evaluate whether it supported generally agreed relationships. While the scope of this study is not that of resolving relationships at the root of the tree of life, this is important to make sure that the genes we selected are informative and do not display obvious paralogy or xenology problems. Analyses of the complete dataset failed to converge under both GTR+G and CAT-GTR+G. Irrespective of that the trees inferred under both models reflect current consensus relatively well. CAT-GTR+G analyses in particular invariably found support for the Eocyte tree, even if with *Koarchaeum cryptofilum* as the sister of Eukaryota rather than the Lokiarchaeota (Figure S3.1). Differently, GTR+G analyses found support for either the Eocyte tree (still with *Koarchaeum cryptofilum* as sister of the Eukaryota) or for Woese's Three Domains Tree (Figure S3.2a and S3.2b). RogueNaRok[140] identified five rogue taxa in the dataset (*Koarchaeum cryptofilum, Treponema_pallidum, Fibrobacter succinogenes, Cyanophora paradoxa* and *Actinomadura madurae*). CAT-GTR+G analyses performed after excluding these taxa still failed to converge (Figure S3.3). However, with the exclusion of the relationships among the eukaryotic supergroups, all key relationships in the CAT-GTR+G tree of Figure S3.3 are resolved according to common knowledge. The GTR+G analysis of the RogueNaRok reduced dataset (Figure S3.4), converged well and resolved the tree in essential agreement with the CAT-GTR+G analysis, supporting in particular the Lokiarchaeota as the sister of the Eukaryota. Overall, these results indicate that instability is limited to the tip-ward part of the tree and this is not unsurprising given that we specifically targeted highly conserved genes to better date the history of life closer to the root rather than the tips. The only area in which our converged GTR+G tree, and our unconverged CAT-GTR+G, tree disagreed with the current consensus were the relationships of the eukaryotic supergroups. This might indicate Long Branch Attraction Artifacts. To test this hypothesis we performed a CAT-GTR+G analysis including

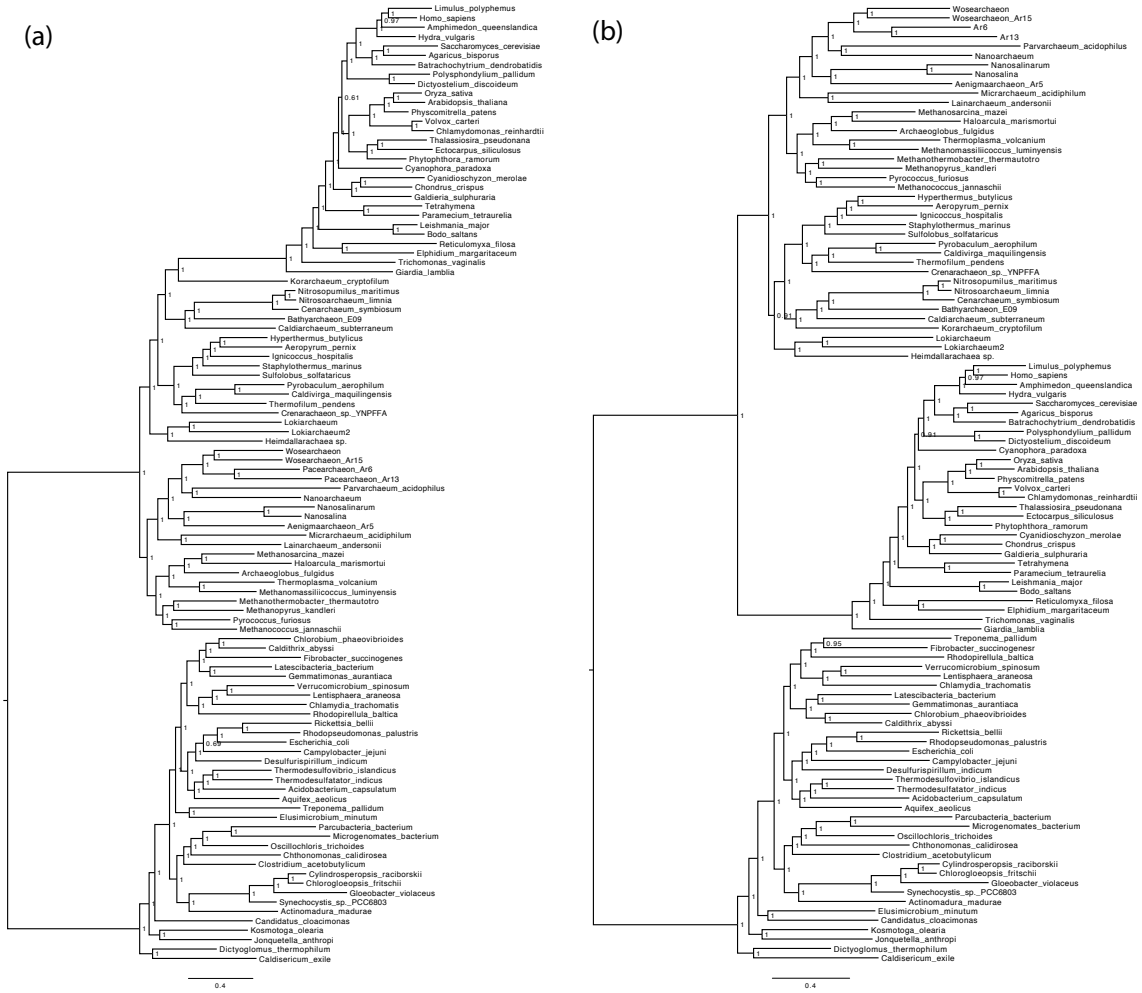
526 only the eukaryotic taxa and found relationships that are fully compatible with the current
527 consensus (Figure S3.5). This indicates that the eukaryotic relationships in Figure S3.3 and
528 S3.4 probably represent tree reconstruction artefacts caused by the attraction between
529 eukaryotes lineages (like the secondarily amitocondriate *Giardia lamblia*) and the prokaryotes.
530 Accordingly, for our clock analyses we used a fixed tree topology compatible with the trees in
531 Figure S3.3 and S3.4, but where the eukaryotes were resolved as in Figure S3.5 and unstable
532 taxa identified by RogueNaRock[151] were reintroduced and resolved according current
533 consensus. This tree is reported in Figure 3 in the main text.
534



535

536 **S3.1**. Phylogeny produced using PhyloBayes with a CAT-GTR+G model (not converged and

537 including rogue taxa).



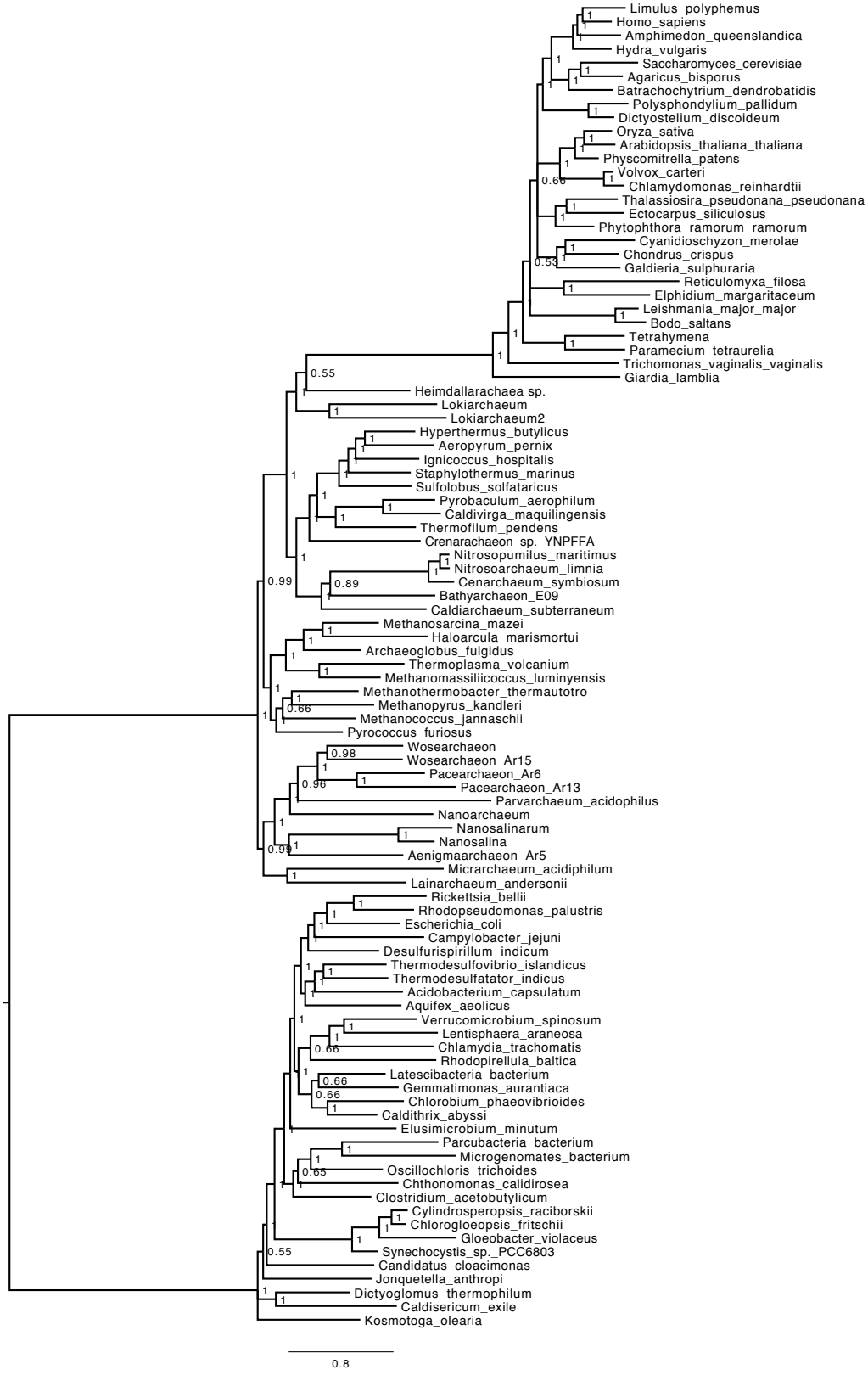
538

539 **S3.2**. Phylogenies produced by two independent runs using PhyloBayes with a GTR+G

540 model (not converged and including rogue taxa) (a) Showing support for the eocyte tree and

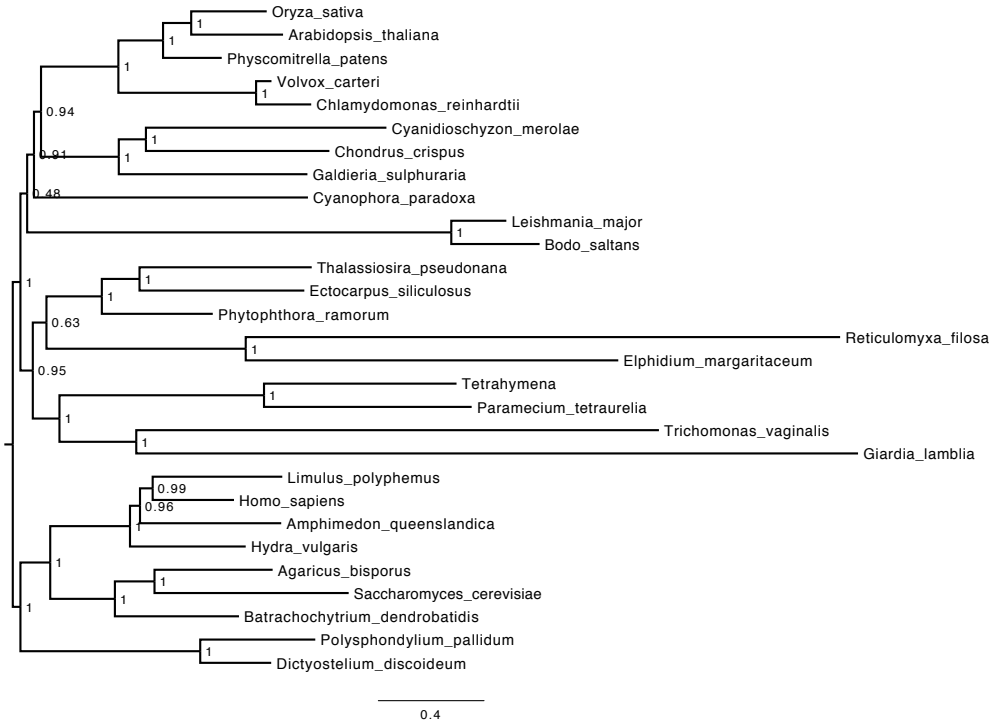
541 (b) for Woese's Three Domains Tree.

542



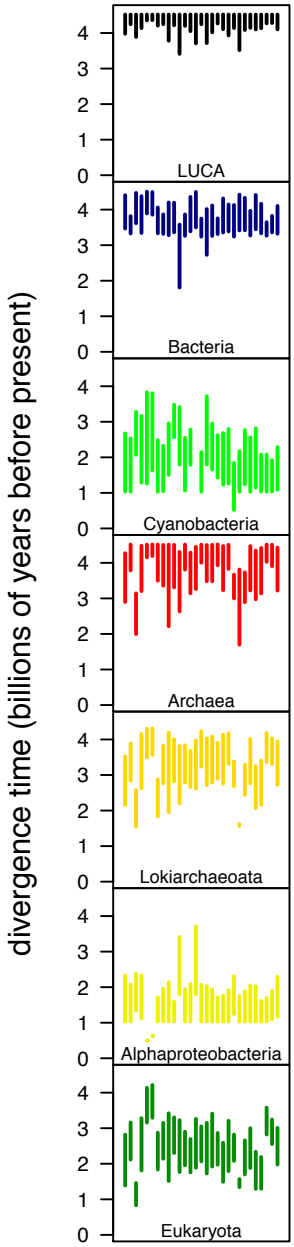
S3.3. Phylogeny produced using PhyloBayes with a CAT-GTR+G model (not converged and excluding rogue taxa).

S3.4. Phylogeny produced using PhyloBayes with a GTR+G model. This analysis converged well (number of cycles = 3872; Burnin = 1000; BPcomp Maxdiff = 0.18; Tracecomp Minimum Effective Size = 244; Tracecomp maximum relative difference = 0.15).

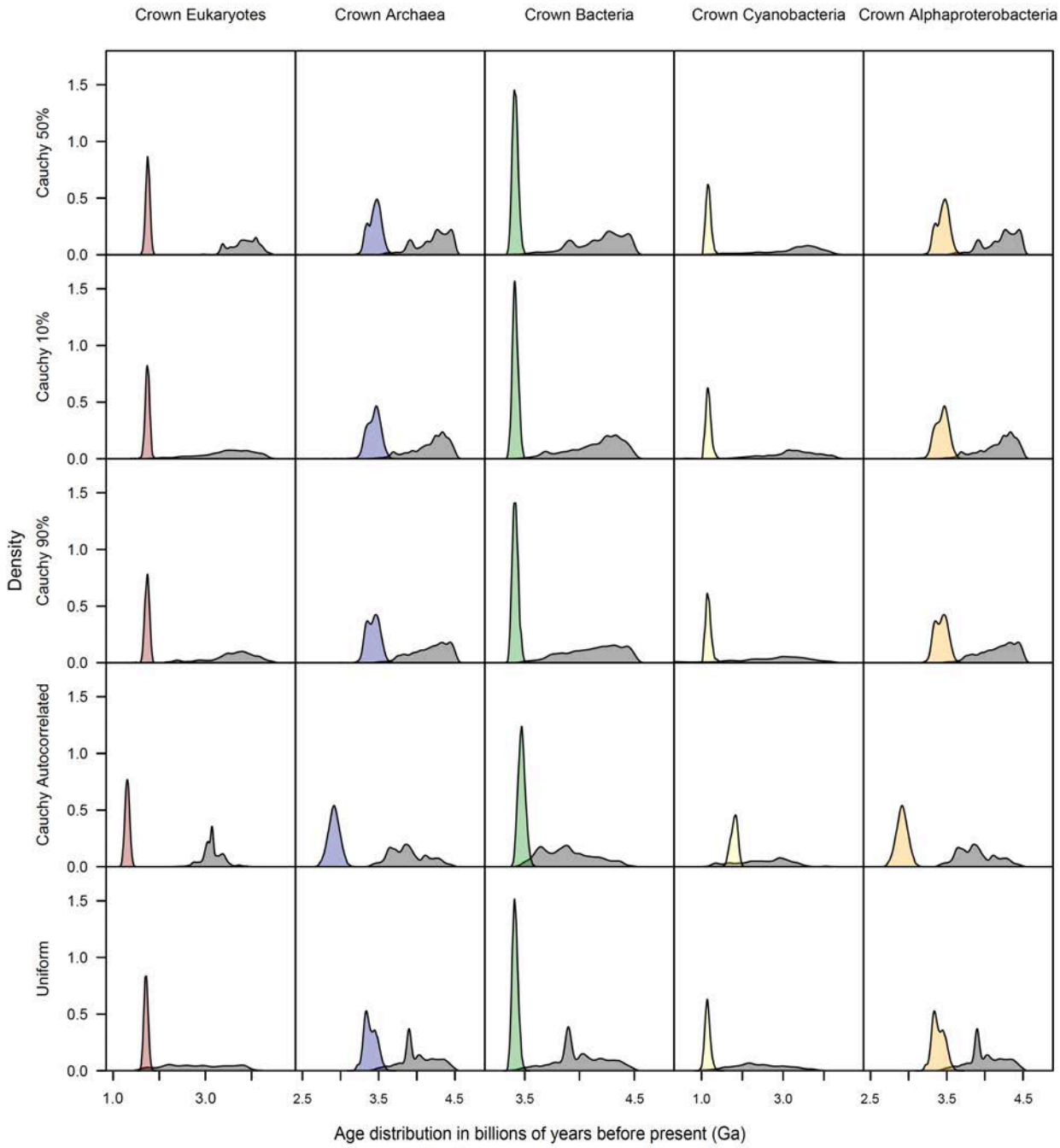


S3.5. Phylogeny showing the Eukaryote only relationships. Produced using PhyloBayes with a CAT-GTR+G model. This analysis reached an acceptable level of convergence (number of cycles = 34660; Burnin = 15000; BPcomp Maxdiff = 0.05; Tracecomp Minimum Effective Size = 170; Tracecomp maximum relative difference = 2.2).

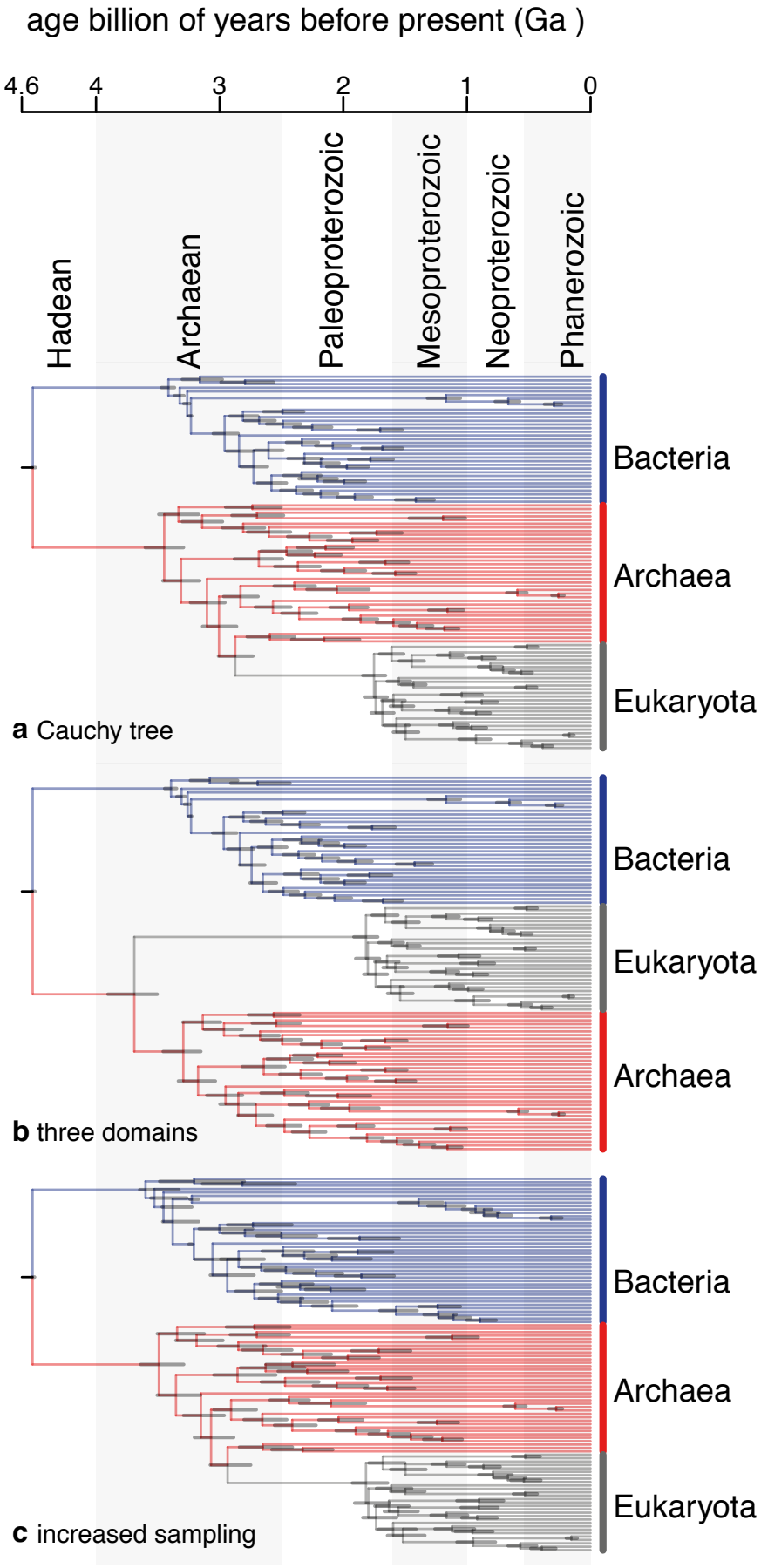
S4 Supplementary figures



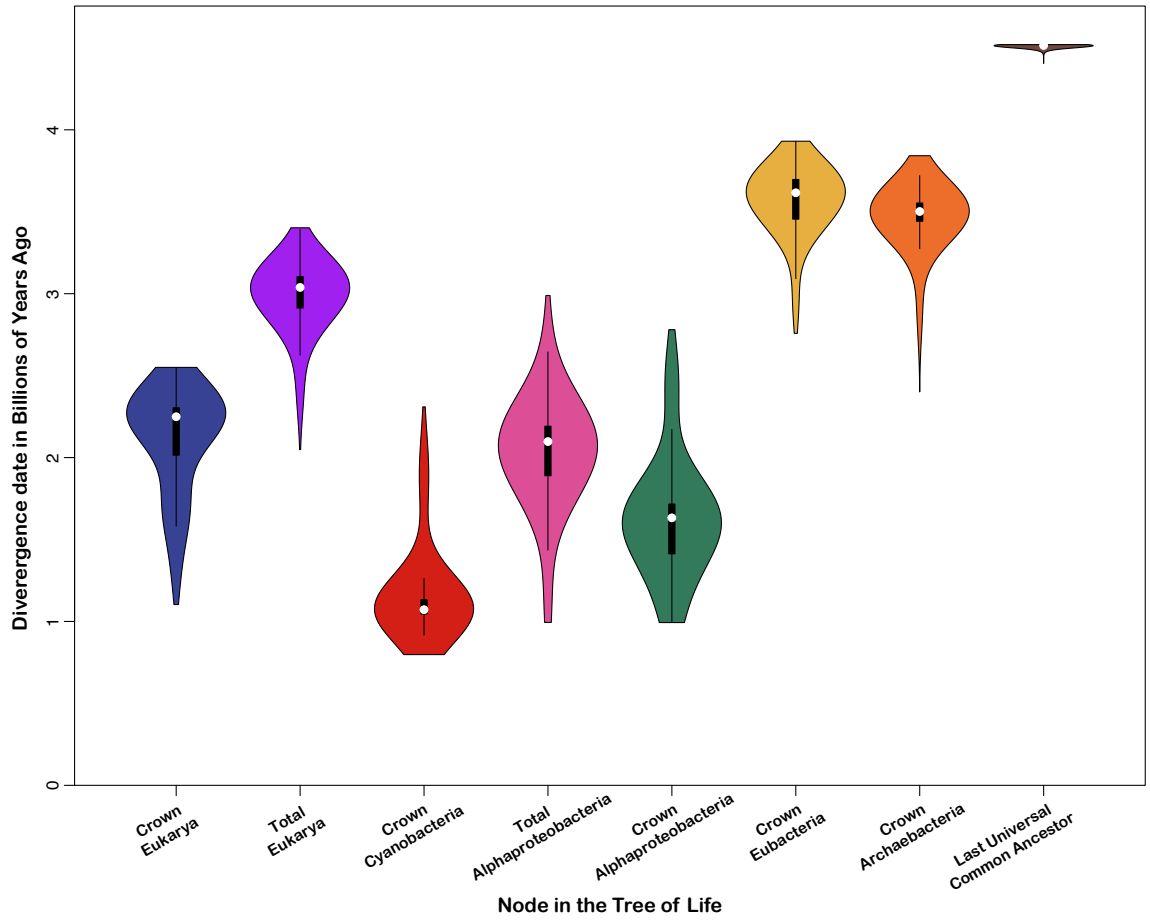
S4.1. Divergence dates for 7 key nodes in the tree of life produced by implementing the molecular clock on a gene by gene basis. In each case a Cauchy 50% calibration distribution density and an uncorrelated clock model was used. On each of the plots the bars represent the divergence dates for genes 1-29.



S4.2. Density plots comparing the prior (grey) and the posterior distributions (colour) in divergence times for 5 nodes in the tree of life. The different calibration density distributions and clock models used are listed along the right side.

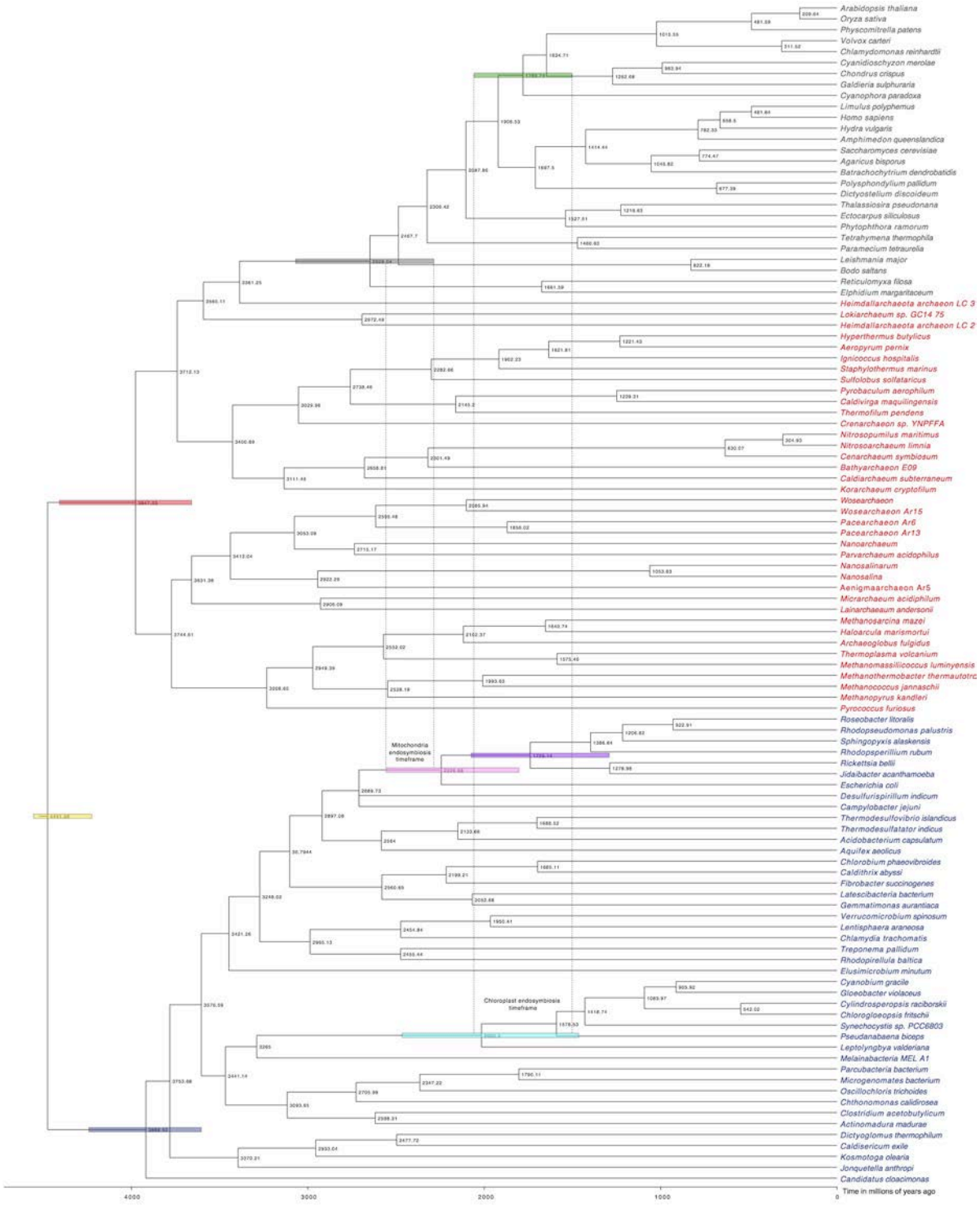

age billion of years before present (Ga)
4.6
4
3
2
1
0
Hadean
Archaean
Paleoproterozoic
Mesoproterozoic
Neoproterozoic
Phanerozoic
Bacteria
Archaea
Eukaryota
a Cauchy tree
Bacteria
Eukaryota
Archaea
b three domains
Bacteria
Archaea
Eukaryota
c increased sampling

S4.3. Comparison of divergence dates produced using (a) a Cauchy 50% calibration distribution density with Eocyte topology (see also Figure 1a), (b) a Cauchy 50% calibration distribution density with a Three Domain Topology, and (c) a Cauchy 50% calibration distribution density with additional species in Alphaproteobacteria and Cyanobacteria. The Eukaryota are highlighted in grey, the Archaebacteria in red and the Eubacteria in blue.



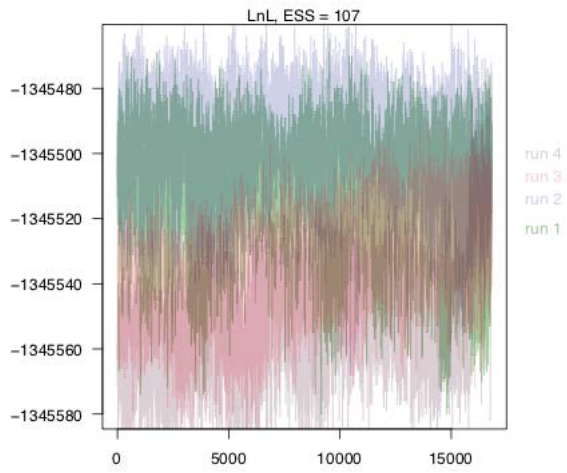
S4.4. Violin plots showing the spread of divergence dates for key nodes in the tree of life from 20 different analyses: Cauchy 50% calibration distribution density; Cauchy 10% calibration distribution density; Cauchy 90% calibration distribution density; Cauchy 50% calibration distribution density with an autocorrelated clock model; Uniform calibration distribution

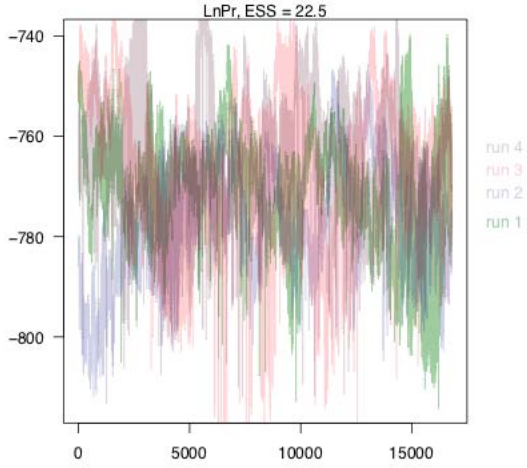
density; and the 15 tree topologies in the 95% credible set of trees from our original phylobayes analysis.



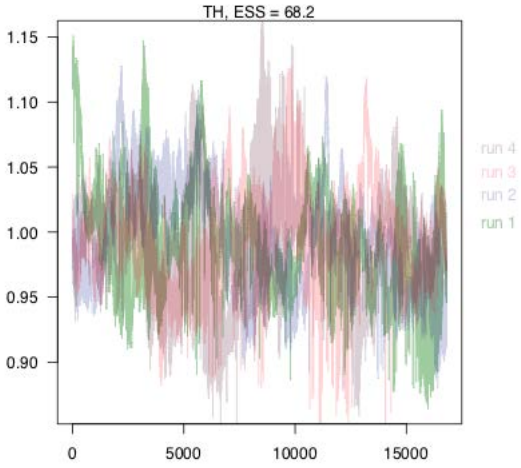
S4.5 Results obtained from an attempt at co-estimating time and topology (20,000,000 generations). Following the advice of one of the reviewers we attempted completing a co-estimation of time and topology using our dataset. This analysis was run in MrBayes 3.2.6[152]

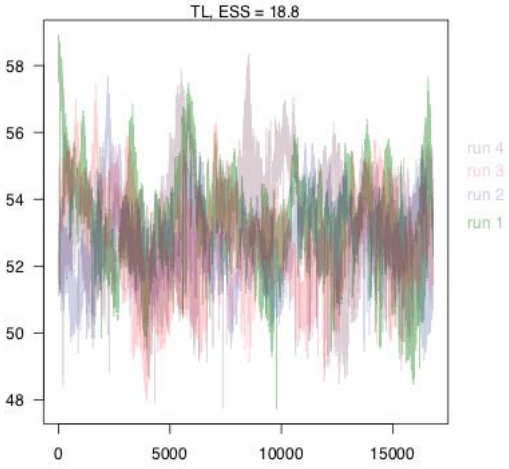
602 under the LG model of substitution with a discrete gamma model of rate variation with four
603 bins. A uniform prior was placed on the topology, except for the 10 internal nodes with set
604 time priors which were constrained to be monophyletic. Prior time constraints on these nodes
605 and the root were set as uniform distributions with the bounds taken from the fossil ages – as
606 in all our other analyses. Branch rates were sampled assuming an uncorrelated Independent
607 Gamma Rates (IGR) model[153] with variance sampled from an exponential distribution (mean
608 = 10). The MCMC model sampled every 1000 generations with four independent runs. The
609 tree was summarised as a 50% majority-rule consensus, and model convergence was assessed
610 by analysing Potential Scale Reduction Factor (PSRF, target < 1.05), Effective Sample Size
611 (ESS, target > 200), and visual inspection using TRACER.
612 Although the results we obtained using co-estimation of time and topology are
613 consistent with those of our other analyses, the co-estimation MCMC runs did not converge
614 within a reasonable amount of computational time (20,000,000 generations), and so they
615 cannot be used to draw definitive conclusions. The similarity between the MCMC samples
616 drawn under co-estimation and those of our other analyses - particularly the well-converged
617 analysis in which we dated the 95% credibility set of topologies (S4.4) - suggest that, at least
618 in this case, there may be little practical advantage in joint estimation when compared to two-
619 step analysis[154,155].
620


LnL, ESS = 107
−1345480
−1345500
−1345520
−1345540
−1345560
−1345580
0
5000
10000
15000
run 4
run 3
run 2
run 1

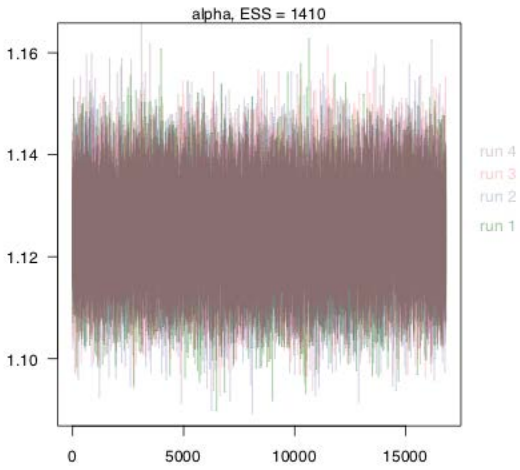

LnPr, ESS = 22.5
−740
−760
−780
−800
0
5000
10000
15000
run 4
run 3
run 2
run 1

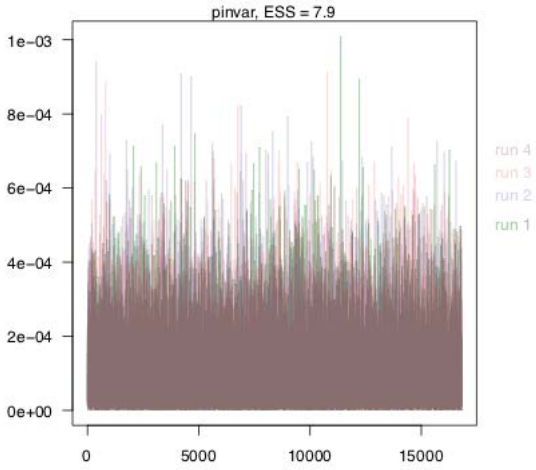
621


TH, ESS = 68.2
1.15
1.10
1.05
1.00
0.95
0.90
0
5000
10000
15000
run 4
run 3
run 2
run 1

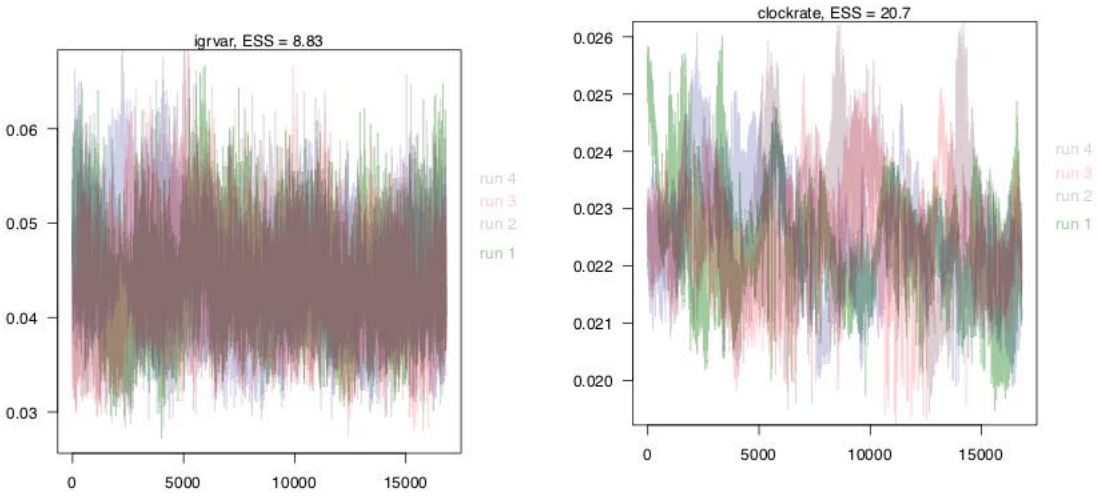

TL, ESS = 18.8
58
56
54
52
50
48
0
5000
10000
15000
run 4
run 3
run 2
run 1

622

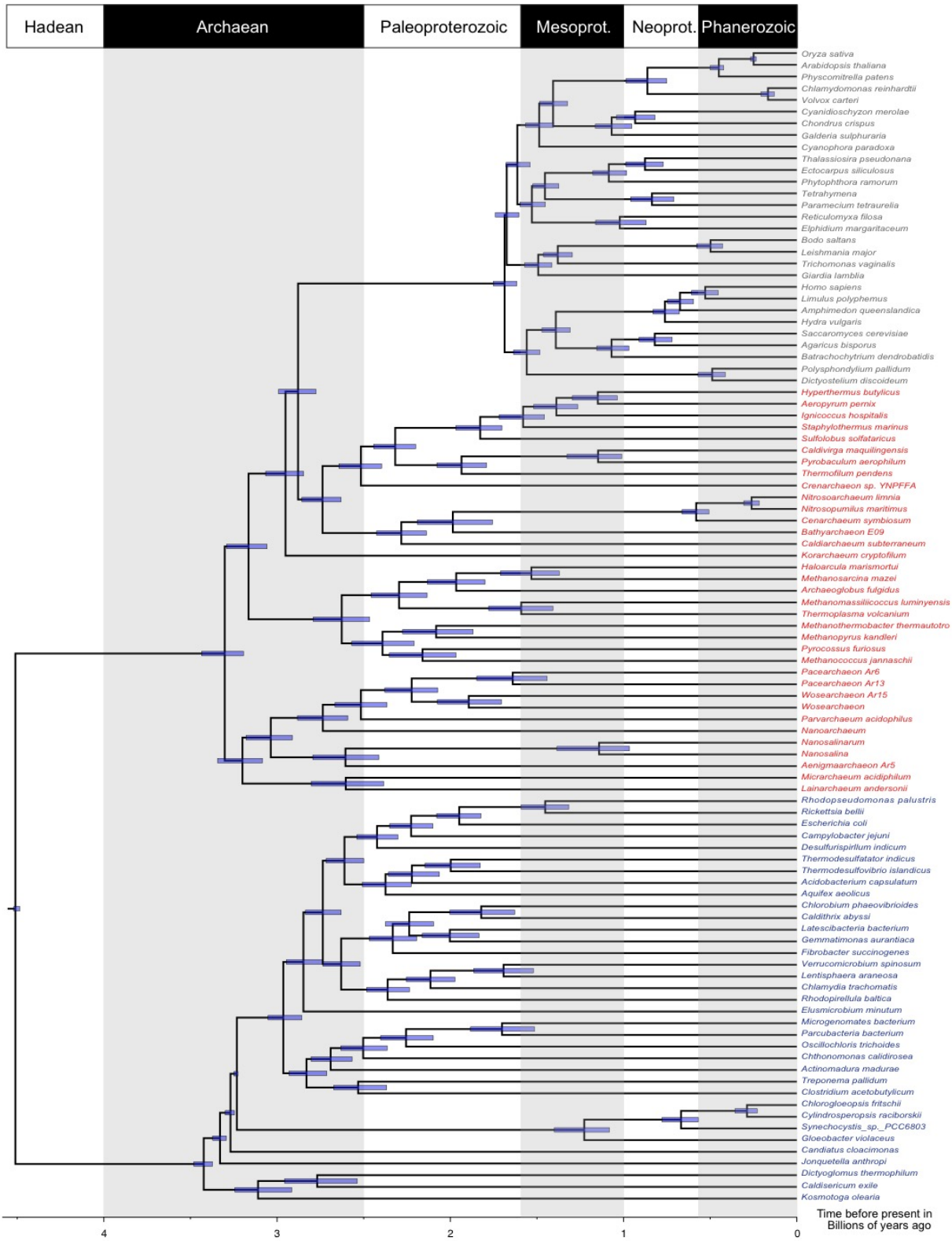

alpha, ESS = 1410
1.16
1.14
1.12
1.10
0
5000
10000
15000
run 4
run 3
run 2
run 1


pinvar, ESS = 7.9
1e−03
8e−04
6e−04
4e−04
2e−04
0e+00
0
5000
10000
15000
run 4
run 3
run 2
run 1

623



S4.6 Convergence statistics for the co-estimation of time and topology analyses. Traces and ESS (after 20,000,000) clearly indicate that the analysis is still far from convergence.



S4.7. Divergence times produced using a Cauchy 50% calibration density distribution and an uncorrelated clock model with the Asgardarchaeota removed. The Eukaryota are highlighted in grey, the Archaebacteria in red and the Eubacteria in blue.

1 Nelson, D. R. 178042: altered volcaniclastic sandstone, Table Top Well; Geochronology dataset 564; *in* Compilation of geochronology data, June 2007 update: Geological Survey of Western Australia. (2005).

2 Hanan, B. B. & Tilton, G. R. 60025: relict of primitive lunar crust? *Earth and Planetary Science Letters* **84**, 15-21 (1987).

3 Barboni, M. *et al.* Early formation of the Moon 4.51 billion years ago. *Science Advances* **3** (2017).

4 Pflug, H. D. & Jaeschke-Boyer, H. Combined structural and chemical analysis of 3,800-Myr-old microfossils. *Nature* **280**, 483-486 (1979).

5 Dodd, M. S. *et al.* Evidence for early life in Earth's oldest hydrothermal vent precipitates. *Nature* **543**, 60-64 (2017).

6 Nutman, A. P., Bennett, V. C., Friend, C. R. L., Van Kranendonk, M. J. & Chivas, A. R. Rapid emergence of life shown by discovery of 3,700-million-year-old microbial structures. *Nature* **537**, 535-538 (2016).

7 Rosing, M. T. ^{13}C-Depleted Carbon Microparticles in >3700-Ma Sea-Floor Sedimentary Rocks from West Greenland. *Science* **283**, 674-676 (1999).

8 Mojzsis, S. J. *et al.* Evidence for life on Earth before 3,800 million years ago. *Nature* **384**, 55 (1996).

9 Schidlowski, M. A 3,800-million-year isotopic record of life from carbon in sedimentary rocks. *Nature* **333**, 313-318 (1988).

10 van Zuilen, M. A., Lepland, A. & Arrhenius, G. Reassessing the evidence for the earliest traces of life. *Nature* **418**, 627-630 (2002).

11 van Zuilen, M. A. *et al.* Graphite and carbonates in the 3.8 Ga old Isua Supracrustal Belt, southern West Greenland. *Precambrian Research* **126**, 331-348 (2003).

12 Lepland, A., Arrhenius, G. & Cornell, D. Apatite in early Archean Isua supracrustal rocks, southern West Greenland: its origin, association with graphite and potential as a biomarker. *Precambrian Research* **118**, 221-241 (2002).

13 Horita, J. & Berndt, M. E. Abiogenic Methane Formation and Isotopic Fractionation Under Hydrothermal Conditions. *Science* **285**, 1055-1057 (1999).

14 Sherwood Lollar, B., Westgate, T. D., Ward, J. A., Slater, G. F. & Lacrampe-Couloume, G. Abiogenic formation of alkanes in the Earth's crust as a minor source for global hydrocarbon reservoirs. *Nature* **416**, 522-524 (2002).

15 Shen, Y., Buick, R. & Canfield, D. E. Isotopic evidence for microbial sulphate reduction in the early Archaean era. *Nature* **410**, 77-81 (2001).

16 Schopf, J. W. Microfossils of the Early Archean Apex Chert: New Evidence of the Antiquity of Life. *Science* **260**, 640-646 (1993).

17 Buick, R. Microfossil Recognition in Archean Rocks: An Appraisal of Spheroids and Filaments from a 3500 M.Y. Old Chert-Barite Unit at North Pole, Western Australia. *PALAIOS* **5**, 441-459 (1990).

18 Ueno, Y., Maruyama, S., Isozaki, Y. & Yurimoto, H. Early Archean (ca. 3.5 Ga) microfossils and 13C-depleted carbonaceous matter in the North Pole area, Western Australia: Field occurrence and geochemistry. *Geochemistry and the Origin of Life*, 203-236 (2001).

19 Brasier, M. D. *et al.* Questioning the evidence for Earth's oldest fossils. *Nature* **416**, 76-81 (2002).

20 Brasier, M. D. *et al.* Critical testing of Earth's oldest putative fossil assemblage from the ~3.5 Ga Apex chert, Chinaman Creek, Western Australia. *Precambrian Research* **140**, 55-102 (2005).

21 Runnegar, B., Dollase, W. A., Ketcham, R. A., Colbert, M. & Carlson, W. D. in *Geol. Soc. Am. Abstracts with Programs.*

22 Engel, A. E. J. *et al.* Alga-Like Forms in Onverwacht Series, South Africa: Oldest Recognized Lifelike Forms on Earth. *Science* **161**, 1005-1008 (1968).

23 Walsh, M. M. & Lowe, D. R. Filamentous microfossils from the 3,500-Myr-old Onverwacht Group, Barberton Mountain Land, South Africa. *Nature* **314**, 530-532 (1985).

24 Westall, F. *et al.* Early Archean fossil bacteria and biofilms in hydrothermally-influenced sediments from the Barberton greenstone belt, South Africa. *Precambrian Research* **106**, 93-116 (2001).

25 Westall, F. *et al.* Implications of a 3.472–3.333 Gyr-old subaerial microbial mat from the Barberton greenstone belt, South Africa for the UV environmental conditions on the early Earth. *Philosophical Transactions of the Royal Society B: Biological Sciences* **361**, 1857-1876 (2006).

26 Allwood, A. C., Walter, M. R., Kamber, B. S., Marshall, C. P. & Burch, I. W. Stromatolite reef from the Early Archaean era of Australia. *Nature* **441**, 714-718 (2006).

27 Allwood, A. C., Walter, M. R., Burch, I. W. & Kamber, B. S. 3.43 billion-year-old stromatolite reef from the Pilbara Craton of Western Australia: Ecosystem-scale insights to early life on Earth. *Precambrian Research* **158**, 198-227 (2007).

28 Byerly, G. R., Lower, D. R. & Walsh, M. M. Stromatolites from the 3,300-3,500-Myr Swaziland Supergroup, Barberton Mountain Land, South Africa. *Nature* **319**, 489-491 (1986).

29 Hofmann, H. J., Grey, K., Hickman, A. H. & Thorpe, R. I. Origin of 3.45 Ga coniform stromatolites in Warrawoona Group, Western Australia. *Geological Society of America Bulletin* **111**, 1256-1262 (1999).

30 Walter, M. R., Buick, R. & Dunlop, J. S. R. Stromatolites 3,400-3,500 Myr old from the North Pole area, Western Australia. *Nature* **284**, 443-445 (1980).

31 Van Kranendonk, M. J. Volcanic degassing, hydrothermal circulation and the flourishing of early life on Earth: A review of the evidence from c. 3490-3240 Ma rocks of the Pilbara Supergroup, Pilbara Craton, Western Australia. *Earth-Science Reviews* **74**, 197-240 (2006).

32 McLoughlin, N., Wilson, L. A. & Brasier, M. D. Growth of synthetic stromatolites and wrinkle structures in the absence of microbes – implications for the early fossil record. *Geobiology* **6**, 95-105 (2008).

33 Lowe, D. R. Abiological origin of described stromatolites older than 3.2 Ga. *Geology* **22**, 387-390 (1994).

34 Javaux, E. J., Marshall, C. P. & Bekker, A. Organic-walled microfossils in 3.2-billion-year-old shallow-marine siliciclastic deposits. *Nature* **463**, 934-938 (2010).

35 Dworkin, M., Falkow, S., Rosenberg, E., Schleifer, K.-H. & Stackebrandt, E. *The Prokaryotes*. 3rd edn edn, 1156-1163, 31-115 (Springer, 2006).

36 Sugitani, K. *et al.* Early evolution of large micro-organisms with cytological complexity revealed by microanalyses of 3.4 Ga organic-walled microfossils. *Geobiology* **13**, 507-521 (2015).

37 Sugitani, K. *et al.* Biogenicity of morphologically diverse carbonaceous microstructures from the ca. 3400 Ma Strelley Pool Formation, in the Pilbara Craton, Western Australia. *Astrobiology* **10**, 899-920 (2010).

38 Sugitani, K., Mimura, K., Nagaoka, T., Lepot, K. & Takeuchi, M. Microfossil assemblage from the 3400 Ma Strelley Pool Formation in the Pilbara Craton, Western Australia: Results form a new locality. *Precambrian Research* **226**, 59-74 (2013).

39 Wacey, D., Kilburn, M. R., Saunders, M., Cliff, J. & Brasier, M. D. Microfossils of sulphur-metabolizing cells in 3.4-billion-year-old rocks of Western Australia. *Nature Geosci* **4**, 698-702 (2011).

40 Wacey, D., McLoughlin, N., Whitehouse, M. J. & Kilburn, M. R. Two coexisting sulfur metabolisms in a ca. 3400 Ma sandstone. *Geology* **38**, 1115-1118 (2010).

41 Duda, J.-P. *et al.* A Rare Glimpse of Paleoarchean Life: Geobiology of an Exceptionally Preserved Microbial Mat Facies from the 3.4 Ga Strelley Pool Formation, Western Australia. *PLOS ONE* **11**, e0147629, (2016).

42 Wacey, D. Stromatolites in the~ 3400 Ma Strelley Pool Formation, Western Australia: examining biogenicity from the macro-to the nano-scale. *Astrobiology* **10**, 381-395 (2010).

43 Lepot, K. *et al.* Texture-specific isotopic compositions in 3.4 Gyr old organic matter support selective preservation in cell-like structures. *Geochimica et Cosmochimica Acta* **112**, 66-86 (2013).

44 Sugitani, K. *et al.* A Paleoarchean coastal hydrothermal field inhabited by diverse microbial communities: the Strelley Pool Formation, Pilbara Craton, Western Australia. *Geobiology* **13**, 522-545, (2015).

45 Hickman, A. Regional review of the 3426–3350 Ma Strelley Pool Formation, Pilbara Craton, Western Australia. *West Australia Geolog Surv Rec* **2008**, 15 (2008).

753 46 Ryder, G. Mass flux in the ancient Earth-Moon system and benign implications for
754 the origin of life on Earth. *Journal of Geophysical Research: Planets* **107**, 6-1-6-13
755 (2002).
756 47 Valley, J. W., Peck, W. H., King, E. M. & Wilde, S. A. A cool early Earth. *Geology*
757 **30**, 351-354 (2002).
758 48 Abramov, O. & Mojzsis, S. J. Microbial habitability of the Hadean Earth during the
759 late heavy bombardment. *Nature* **459**, 419-422 (2009).
760 49 Koeberl, C. Impact processes on the early Earth. *Elements* **2**, 211-216 (2006).
761 50 Kleine, T., Palme, H., Mezger, K. & Halliday, A. N. Hf-W Chronometry of Lunar
762 Metals and the Age and Early Differentiation of the Moon. *Science* **310**, 1671-1674
763 (2005).
764 51 Carlson, R. W. & Lugmair, G. W. The age of ferroan anorthosite 60025: oldest crust
765 on a young Moon? *Earth and Planetary Science Letters* **90**, 119-130 (1988).
766 52 Bottke, W. F. *et al.* Dating the Moon-forming impact event with asteroidal meteorites.
767 *Science* **348**, 321-323 (2015).
768 53 Jacobson, S. A. *et al.* Highly siderophile elements in Earth/'s mantle as a clock for the
769 Moon-forming impact. *Nature* **508**, 84-87 (2014).
770 54 Tera, F., Papanastassiou, D. & Wasserburg, G. in *Lunar and Planetary Science*
771 *Conference.*
772 55 Halliday, A., Rehkämper, M., Lee, D.-C. & Yi, W. Early evolution of the Earth and
773 Moon: new constraints from Hf-W isotope geochemistry. *Earth and Planetary*
774 *Science Letters* **142**, 75-89 (1996).
775 56 Halliday, A. N. A young Moon-forming giant impact at 70–110 million years
776 accompanied by late-stage mixing, core formation and degassing of the Earth.

Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences **366**, 4163-4181 (2008).

57 Kleine, T., Munker, C., Mezger, K. & Palme, H. Rapid accretion and early core formation on asteroids and the terrestrial planets from Hf-W chronometry. *Nature* **418**, 952-955 (2002).

58 Touboul, M., Kleine, T., Bourdon, B., Palme, H. & Wieler, R. Late formation and prolonged differentiation of the Moon inferred from W isotopes in lunar metals. *Nature* **450**, 1206-1209 (2007).

59 Halliday, A. The origin and earliest history of the Earth. *Planets, Asteriods, Comets and The Solar System*, 149-211 (2014).

60 Kamo, S. L. & Davis, D. W. Reassessment of Archean crustal development in the Barberton Mountain Land, South Africa, based on U-Pb dating. *Tectonics* **13**, 167-192 (1994).

61 Cairns-Smith, A. G. Precambrian solution photochemistry, inverse segregation, and banded iron formations. *Nature* **276**, 807-808 (1978).

62 Crowe, S. A. *et al.* Photoferrotrophs thrive in an Archean Ocean analogue. *Proceedings of the National Academy of Sciences* **105**, 15938-15943 (2008).

63 Konhauser, K. O. *et al.* Could bacteria have formed the Precambrian banded iron formations? *Geology* **30**, 1079-1082 (2002).

64 Grotzinger, J. P. & Rothman, D. H. An abiotic model for stromatolite morphogenesis. *Nature* **383**, 423-425 (1996).

65 Satkoski, A. M., Beukes, N. J., Li, W., Beard, B. L. & Johnson, C. M. A redox-stratified ocean 3.2 billion years ago. *Earth and Planetary Science Letters* **430**, 43-53 (2015).

66 Olson, S. L., Kump, L. R. & Kasting, J. F. Quantifying the areal extent and dissolved oxygen concentrations of Archean oxygen oases. *Chemical Geology* **362**, 35-43 (2013).

67 Homann, M., Heubeck, C., Airo, A. & Tice, M. M. Morphological adaptations of 3.22 Ga-old tufted microbial mats to Archean coastal habitats (Moodies Group, Barberton Greenstone Belt, South Africa). *Precambrian Research* **266**, 47-64 (2015).

68 Anbar, A. D. *et al.* A Whiff of Oxygen Before the Great Oxidation Event? *Science* **317**, 1903-1906 (2007).

69 Crowe, S. A. *et al.* Atmospheric oxygenation three billion years ago. *Nature* **501**, 535-538 (2013).

70 Kendall, B. *et al.* Pervasive oxygenation along late Archaean ocean margins. *Nature Geosci* **3**, 647-652 (2010).

71 Czaja, A. D. *et al.* Evidence for free oxygen in the Neoarchean ocean based on coupled iron–molybdenum isotope fractionation. *Geochimica et Cosmochimica Acta* **86**, 118-137 (2012).

72 Planavsky, N. J. *et al.* Evidence for oxygenic photosynthesis half a billion years before the Great Oxidation Event. *Nature Geosci* **7**, 283-286 (2014).

73 Riding, R., Fralick, P. & Liang, L. Identification of an Archean marine oxygen oasis. *Precambrian Research* **251**, 232-237 (2014).

74 Byerly, G. R., Kröner, A., Lowe, D. R., Todt, W. & Walsh, M. M. Prolonged magmatism and time constraints for sediment deposition in the early Archean Barberton greenstone belt: evidence from the Upper Onverwacht and Fig Tree groups. *Precambrian Research* **78**, 125-138 (1996).

75 Li, H. *et al.* Recent advances in the study of the Mesoproterozoic geochronology in the North China Craton. *Journal of Asian Earth Sciences* 72, 216-227, (2013).

826 76 Peng, Y., Bao, H. & Yuan, X. New morphological observations for Paleoproterozoic
827 acritarchs from the Chuanlinggou Formation, North China. *Precambrian Research*
828 **168**, 223-232 (2009).
829 77 Lamb, D. M., Awramik, S. M., Chapman, D. J. & Zhu, S. Evidence for eukaryotic
830 diversification in the~ 1800 million-year-old Changzhougou Formation, North China.
831 *Precambrian Research* **173**, 93-104 (2009).
832 78 MoczydŁOwska, M., Landing, E. D., Zang, W. & Palacios, T. Proterozoic
833 phytoplankton and timing of Chlorophyte algae origins. *Palaeontology* **54**, 721-733
834 (2011).
835 79 Zhu, S. *et al.* Decimetre-scale multicellular eukaryotes from the 1.56-billion-year-old
836 Gaoyuzhuang Formation in North China. *Nature Communications* **7**, 11500 (2016).
837 80 Horodyski, R. J. Problematic Bedding-Plane Markings from the Middle Proterozoic
838 Appekunny Argillite, Belt Supergroup, Northwestern Montana. *Journal of*
839 *Paleontology* **56**, 882-889 (1982).
840 81 Yin, L.-m. Acanthomorphic acritarchs from Meso-Neoproterozoic shales of the
841 Ruyang Group, Shanxi, China. *Review of Palaeobotany and Palynology* **98**, 15-25
842 (1997).
843 82 Knoll, A. H., Javaux, E. J., Hewitt, D. & Cohen, P. Eukaryotic organisms in
844 Proterozoic oceans. *Philosophical Transactions of the Royal Society B: Biological*
845 *Sciences* **361**, 1023-1038 (2006).
846 83 Cohen, P. A. & Macdonald, F. A. The Proterozoic Record of Eukaryotes.
847 *Paleobiology* **41**, 610-632 (2015).
848 84 Knoll, A. H. Paleobiological perspectives on early eukaryotic evolution. *Cold Spring*
849 *Harbor Perspectives in Biology* **6**, a016121 (2014).

85 Lu, S., Yang, C. & Zhu, S. in *Proceedings of the 30th International Geological Congress, Geological Publishing House, Beijing.* 13-T105.

86 Turner, E. C. & Kamber, B. S. Arctic Bay Formation, Borden Basin, Nunavut (Canada): Basin evolution, black shale, and dissolved metal systematics in the Mesoproterozoic ocean. *Precambrian Research* **208–211**, 1-18 (2012).

87 Bengston, S. *et al.,* Three-dimensional preservation of cellular and subcellular structures suggests 1.6 billion-year-old crown-group red algae. *PLOS Biology*, **15**, e2000735, (2017).

88 Butterfield, N. J., Knoll, A. H. & Swett, K. A bangiophyte red alga from the Proterozoic of arctic Canada. *Science* **250**, 104-108 (1990).

89 Butterfield, N. J. Bangiomorpha pubescens n. gen., n. sp.: implications for the evolution of sex, multicellularity, and the Mesoproterozoic/Neoproterozoic radiation of eukaryotes. *Paleobiology* **26** (2000).

90 Parfrey, L. W., Lahr, D. J. G., Knoll, A. H. & Katz, L. A. Estimating the timing of early eukaryotic diversification with multigene molecular clocks. *Proceedings of the National Academy of Sciences* **108**, 13624-13629 (2011).

91 Eme, L., Sharpe, S. C., Brown, M. W. & Roger, A. J. On the age of eukaryotes: evaluating evidence from fossils and molecular clocks. *Cold Spring Harb. Perspect. Biol.* **6**, a016139 (2014).

92 Hoek, C., Mann, D. & Jahns, H. M. *Algae: an introduction to phycology*. (Cambridge university press, 1995).

93 Lee, R. E. *Phycology*. (Cambridge University Press, 2008).

94 Yang, E. C. *et al.* Divergence time estimates and the evolution of major lineages in the florideophyte red algae. *Scientific Reports* **6**, 21361 (2016).

95 Heaman, L., LeCheminant, A. & Rainbird, R. in *Geological Association of Canada, Programs with Abstracts.* A55.

96 LeCheminant, A. N. & Heaman, L. M. Mackenzie igneous events, Canada: Middle Proterozoic hotspot magmatism associated with ocean opening. *Earth and Planetary Science Letters* **96**, 38-48 (1989).

97 Kah, L. C., Sherman, A. G., Narbonne, G. M., Knoll, A. H. & Kaufman, A. J. δ13C stratigraphy of the Proterozoic Bylot Supergroup, Baffin Island, Canada: implications for regional lithostratigraphic correlations. *Canadian Journal of Earth Sciences* **36**, 313-332 (1999).

98 Mayr, U. *Geology of Eastern Prince of Wales Island and Adjacent Smaller Islands, Nunavut (parts of NTS 68D, Baring Channel and 68A, Fisher Lake)*. Vol. 574 (Geological Survey of Canada, 2004).

99 Long, D. G. F. & Turner, E. C. Tectonic, sedimentary and metallogenic re-evaluation of basal strata in the Mesoproterozoic Bylot basins, Nunavut, Canada: Are unconformity-type uranium concentrations a realistic expectation? *Precambrian Research* **214–215**, 192-209 (2012).

100 Zhongying, Z. Clastic facies microfossils from the Chuanlinggou Formation (1800 Ma) near Jixian, North China. *Journal of Micropalaeontology* **5**, 9-16 (1986).

101 Roger, A. J., Muñoz-Gómez, S. A. & Kamikawa, R. The Origin and Diversification of Mitochondria. *Current Biology* **27**, R1177-R1192, (2017).

102 Williams, K. P., Sobral, B. W. & Dickerman, A. W. A robust species tree for the alphaproteobacteria. *Journal of bacteriology* **189**, 4578-4586 (2007).

103 Wang, Z. & Wu, M. An integrated phylogenomic approach toward pinpointing the origin of mitochondria. *Scientific Reports* **5**, 7949 (2015).

104 Atteia, A. *et al.* A Proteomic Survey of Chlamydomonas reinhardtii Mitochondria Sheds New Light on the Metabolic Plasticity of the Organelle and on the Nature of the α-Proteobacterial Mitochondrial Ancestor. *Molecular Biology and Evolution* **26**, 1533-1548 (2009).

105 Esser, C. *et al.* A Genome Phylogeny for Mitochondria Among α-Proteobacteria and a Predominantly Eubacterial Ancestry of Yeast Nuclear Genes. *Molecular Biology and Evolution* **21**, 1643-1660 (2004).

106 Gray, M. W. Mosaic nature of the mitochondrial proteome: Implications for the origin and evolution of mitochondria. *Proceedings of the National Academy of Sciences* **112**, 10133-10138 (2015).

107 Gray, M. W. Mitochondrial evolution. *Cold Spring Harb. Perspec. Biol.* **4**, a011403 (2012).

108 Shih, P. M. & Matzke, N. J. Primary endosymbiosis events date to the later Proterozoic with cross-calibrated phylogenetic dating of duplicated ATPase proteins. *Proceedings of the National Academy of Sciences* **110**, 12355-12360, (2013).

109 Schopf, J. W. Fossil evidence of Archaean life. *Philosophical Transactions of the Royal Society B: Biological Sciences* **361**, 869-885, (2006).

110 Brasier, M., McLoughlin, N., Green, O. & Wacey, D. A fresh look at the fossil evidence for early Archaean cellular life. *Philosophical Transactions of the Royal Society B: Biological Sciences* **361**, 887-902, (2006)

111 Amard, B. & Bertrand-Sarfati, J. Microfossils in 2000 Ma old cherty stromatolites of the Franceville Group, Gabon. *Precambrian Research* **81**, 197-221, (1997).

112 Tomitani, A., Knoll, A. H., Cavanaugh, C. M. & Ohno, T. The evolutionary diversification of cyanobacteria: Molecular–phylogenetic and paleontological

perspectives. *Proceedings of the National Academy of Sciences* **103**, 5442-5447, (2006).

113 Butterfield, N. J. Proterozoic photosynthesis – a critical review. *Palaeontology* **58**, 953-972, (2015).

114 Golubic, S. & Hofmann, H. J. Comparison of Holocene and Mid-Precambrian Entophysalidaceae (Cyanophyta) in Stromatolitic Algal Mats: Cell Division and Degradation. *Journal of Paleontology* **50**, 1074-1082 (1976).

115 Hofmann, H. J. Precambrian Microflora, Belcher Islands, Canada: Significance and Systematics. *Journal of Paleontology* **50**, 1040-1073 (1976).

116 Sánchez-Baracaldo, P., Raven, J. A., Pisani, D. & Knoll, A. H. Early photosynthetic eukaryotes inhabited low-salinity habitats. *Proceedings of the National Academy of Sciences* **114**, E7737-E7745, (2017).

117 Ponce-Toledo, R. I. *et al.* An Early-Branching Freshwater Cyanobacterium at the Origin of Plastids. *Current Biology* **27**, 386-391 (2017).

118 Ochoa de Alda, J. A. G., Esteban, R., Diago, M. L. & Houmard, J. The plastid ancestor originated among one of the major cyanobacterial lineages. *Nature Communications* **5**, 4937 (2014).

119 Deusch, O. *et al.* Genes of Cyanobacterial Origin in Plant Nuclear Genomes Point to a Heterocyst-Forming Plastid Ancestor. *Molecular Biology and Evolution* **25**, 748-761 (2008).

120 Gradstein, F. M., Ogg, J. G., Schmitz, M. & Ogg, G. *The geologic time scale 2012*. (elsevier, 2012).

121 Taylor, T. N., Hass, H. & Kerp, H. The oldest fossil ascomycetes. *Nature* **399**, 648-648 (1999).

122 Taylor, T. N., Hass, H., Kerp, H., Krings, M. & Hanlin, R. T. Perithecial Ascomycetes from the 400 Million Year Old Rhynie Chert: An Example of Ancestral Polymorphism. *Mycologia* **96**, 1403-1419 (2004).

123 Butterfield, N. J. Probable Proterozoic fungi. *Paleobiology* **31**, 165-182 (2005).

124 Butterfield, N. J. Early evolution of the Eukaryota. *Palaeontology* **58**, 5-17 (2015).

125 Yuan, X., Xiao, S. & Taylor, T. N. Lichen-Like Symbiosis 600 Million Years Ago. *Science* **308**, 1017-1020 (2005).

126 Cunningham, J. A. *et al.* Distinguishing geology from biology in the Ediacaran Doushantuo biota relaxes constraints on the timing of the origin of bilaterians. *Proceedings of the Royal Society B: Biological Sciences* **279**, 2369-2376, (2012).

127 Redecker, D., Kodner, R. & Graham, L. E. Glomalean fungi from the Ordovician. *Science* **289**, 1920-1921 (2000).

128 Bengtson, S. *et al.* Fungus-like mycelial fossils in 2.4-billion-year-old vesicular basalt. *Nature Ecology & Evolution* **1**, 0141 (2017).

129 Schumann, G., Manz, W., Reitner, J. & Lustrino, M. Ancient Fungal Life in North Pacific Eocene Oceanic Crust. *Geomicrobiology Journal* **21**, 241-246 (2004).

130 Ivarsson, M., Bengtson, S., Skogby, H., Belivanova, V. & Marone, F. Fungal colonies in open fractures of subseafloor basalt. *Geo-Marine Letters* **33**, 233-243 (2013).

131 Ivarsson, M. *et al.* Fossilized fungi in subseafloor Eocene basalts. *Geology* **40**, 163-166 (2012).

132 Mark, D. F. *et al.* 40Ar/39Ar dating of hydrothermal activity, biota and gold mineralization in the Rhynie hot-spring system, Aberdeenshire, Scotland. *Geochimica et Cosmochimica Acta* **75**, 555-569 (2011).

133 Parry, S., Noble, S., Crowley, Q. & Wellman, C. A high-precision U–Pb age constraint on the Rhynie Chert Konservat-Lagerstätte: time scale and other implications. *Journal of the Geological Society* **168**, 863-872 (2011).

134 Rice, C. & Ashcroft, W. The geology of the northern half of the Rhynie Basin, Aberdeenshire, Scotland. *Transactions of the Royal Society of Edinburgh: Earth Sciences* **94**, 299-308 (2003).

135 Parry, S., Noble, S., Crowley, Q. & Wellman, C. Reply to Discussion on 'A high-precision U–Pb age constraint on the Rhynie Chert Konservat-Lagerstätte: time scale and other implications' Journal, 168, 863–872. *Journal of the Geological Society* **170**, 703-706 (2013).

136 Wellman, C. H. Spore assemblages from the Lower Devonian 'Lower Old Red Sandstone'deposits of the Rhynie outlier, Scotland. *Transactions of the Royal Society of Edinburgh: Earth Sciences* **97**, 167-211 (2006).

137 Bosak, T. *et al.* Possible early foraminiferans in post-Sturtian (716− 635 Ma) cap carbonates. *Geology* **40**, 67-70 (2012).

138 Bosak, T. *et al.* Agglutinated tests in post-Sturtian cap carbonates of Namibia and Mongolia. *Earth and Planetary Science Letters* **308**, 29-40 (2011).

139 Antcliffe, J. B., Gooday, A. J. & Brasier, M. D. Testing the protozoan hypothesis for Ediacaran fossils: a developmental analysis of Palaeopascichnus. *Palaeontology* **54**, 1157-1175 (2011).

140 Culver, S. J. Early Cambrian foraminifera from west Africa. *Science* **254**, 689-692 (1991).

141 McIlroy, D., Green, O. R. & Brasier, M. D. Palaeobiology and evolution of the earliest agglutinated Foraminifera: Platysolenites, Spirosolenites and related forms. *Lethaia* **34**, 13-29 (2001).

142 Villeneuve, M., Theveniaut, H., Ndiaye, P. M. & Retière, S. Re-assessment of the northern Guinean "Koubia–Lessere unconformity"(KLU): Consequences on the geological correlations throughout West Africa. *Comptes Rendus Geoscience* **346**, 262-272 (2014).

143 Rozanov, A. Y. Platysolenites. *Upper Precambrian and Cambrian palaeontology of the East-European platform*, 94-100 (1983).

144 Rozanov, A. Y. & Zhuravlev, A. Y. in *Origin and early evolution of the Metazoa* 205-282 (Springer, 1992).

145 Lipps, J. Proterozoic and Cambrian skeletonized protists. *The Proterozoic Biosphere*, 237-240 (1992).

146 Lipps, J. Origin and early evolution of Foraminifera. *Studies in benthic Foraminifera* **90**, 3-9 (1992).

147 Lipps, J. & Rozanov, A. Y. The late Precambrian-Cambrian agglutinated fossil Platysolenites. *Paleontological Journal of Paleontologicheskii Zhurnal* **30**, 679-687 (1996).

148 Porter, S. M. & Knoll, A. H. Testate amoebae in the Neoproterozoic Era: evidence from vase-shaped microfossils in the Chuar Group, Grand Canyon. *Paleobiology* **26**, 360-385 (2000).

149 Clark, J. W. & Donoghue, P. C. J. Constraining the timing of whole genome duplication in plant evolutionary history. *Proceedings of the Royal Society B: Biological Sciences* **284**, (2017).

150 Benton, M. J. *et al.* Constraints on the timescale of animal evolutionary history. *Palaeontologia Electronica* **18**, 1-106 (2015).

151 Aberer, A. J., Krompass, D. & Stamatakis, A. Pruning Rogue Taxa Improves Phylogenetic Accuracy: An Efficient Algorithm and Webservice. *Systematic Biology* **62**, 162-166 (2013).

152 Ronquist, F., *et al*. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Systematic biology* **61**, 539-542 (2012).

153 Lepage, T., *et al.* A general comparison of relaxed molecular clock models. *Molecular biology and evolution* **24**, 2669-2680 (2007).

154 Rannala, B. Conceptual issues in Bayesian divergence time estimation. *Philos Trans R Soc Lond B Biol Sci*. **371**, 20150134 (2016).

155 Donoghue P.C.J. and Yang Z. The evolution of methods for establishing evolutionary timescales. *Philos Trans R Soc Lond B Biol Sci*. **371**, 20160020 (2016).